Datenerhebung und Schätzung bei sensitiven Merkmalen

Inaugural-Dissertation zur
Erlangung der wirtschaftswissenschaftlichen Doktorwürde
des Fachbereichs Wirtschaftswissenschaften
der Philipps-Universität Marburg

eingereicht von
Heiko Grönitz
Diplom-Mathematiker aus Altenburg

Erstgutachter:	Prof. Dr. Karlheinz Fleischer
Zweitgutachter:	Prof. Dr. Sascha Mölls
Einreichungstermin:	07. März 2013
Prüfungstermin:	15. Mai 2013
Hochschulkennziffer:	1180

Inhaltliche Zusammenführung und Zusammenfassung von vier Aufsätzen zum Thema

"Datenerhebung und Schätzung bei sensitiven Merkmalen"

Heiko Grönitz

Die folgende inhaltliche Zusammenführung und Zusammenfassung bezieht sich auf die Manuskripte

1. Groenitz, H. (2012): A New Privacy-Protecting Survey Design for Multichotomous Sensitive Variables. Metrika, DOI: 10.1007/s00184-012-0406-8.

2. Groenitz, H. (2013a): Using Prior Information in Privacy-Protecting Survey Designs for Categorical Sensitive Variables. Article 1 / 2013 in "Discussion Papers on Statistics and Quantitative Methods", Philipps-University Marburg, Faculty of Business Administration, Department of Statistics.

3. Groenitz, H. (2013b): Applying the Nonrandomized Diagonal Model to Estimate a Sensitive Distribution in Complex Sample Surveys. Accepted in: Journal of Statistical Theory and Practice.

4. Groenitz, H. (2013c): A Covariate Nonrandomized Response Model for Multicategorical Sensitive Variables.

Wenn in einer Umfrage Daten über ein Merkmal X gesammelt werden sollen, geht man typischerweise wie folgt vor: Man wählt zufällig einige Personen aus und fragt jede dieser Personen

"Wie lautet Ihre Ausprägung bei dem Merkmal X?"

Diese Direktbefragung ist allerdings problematisch, sobald X ein sensitives Merkmal wie Einkommen, Steuerhinterziehung, Versicherungsbetrug oder politische Präferenzen ist. Bei direkten Fragen, wie z.B.

"Wie hoch ist Ihr Einkommen?" oder "Haben Sie schon einmal Steuern hinterzogen?"

wird es oft Personen geben, die die Antwort verweigern oder eine Falschantwort geben. Würde man aus den erhaltenen Antworten die Verteilung von X schätzen, so ist daher eine erhebliche Verzerrung zu erwarten. Mit anderen Worten: Die geschätzte Verteilung wird in der Regel stark von der tatsächlichen Verteilung abweichen. Aus diesem Grund benötigt man geschickte Umfragetechniken, die einerseits die Privatsphäre der Befragten schützen, anderseits aber Daten liefern, die Rückschlüsse auf die Verteilung des sensitiven Merkmals zulassen.

Einen Beitrag in diesem Forschungsgebiet leistet der Artikel Groenitz (2012). In diesem Aufsatz wird zunächst ein Umfragedesign, das "Diagonal-Modell" (DM), zur Datenerhebung bei kategorialen, sensitiven Merkmalen vorgeschlagen. Sei also X ein sensitives Merkmal mit möglichen Ausprägungen $1, 2, ..., k$ (die Werte könnten z.B. Einkommensklassen repräsentieren). Für das DM muss ein Hilfsmerkmal W, welches ebenfalls die Werte $1, 2, ..., k$ annehmen kann, eine bekannte Verteilung besitzt und als unabhängig von X angesehen werden kann, festgelegt werden. Dabei muss auch darauf geachtet werden, dass dem Interviewer die Werte der Befragten für W nicht bekannt sind. Ein solches Merkmal W könnte z.B. für $k = 4$ wie folgt aussehen:

$$W = \begin{cases} 1, & \text{falls Geburtstag der Mutter zwischen 01. Jan. und 16. Aug.} \\ 2, & \text{falls Geburtstag der Mutter zwischen 17. Aug. und 01. Okt.} \\ 3, & \text{falls Geburtstag der Mutter zwischen 02. Okt. und 16. Nov.} \\ 4, & \text{falls Geburtstag der Mutter zwischen 17. Nov. und 31. Dez.} \end{cases}$$

Ignoriert man Schaltjahre und unterstellt eine gleichmäßige Verteilung der Geburten auf 365 Tage des Jahres, so ist die Verteilung von W durch

Ausprägung	$W = 1$	$W = 2$	$W = 3$	$W = 4$
Anteil	$\frac{228}{365}$	$\frac{46}{365}$	$\frac{46}{365}$	$\frac{45}{365}$

gegeben. Jeder Befragte wird nun instruiert anhand seiner Ausprägungen für X und W eine Antwort A zu geben. Für $k = 4$ enthält die nachfolgende Tabelle die zu gebende Antwort A in Abhängigkeit von X und W:

X/W	$W = 1$	$W = 2$	$W = 3$	$W = 4$
$X = 1$	**1**	**2**	**3**	**4**
$X = 2$	**4**	**1**	**2**	**3**
$X = 3$	**3**	**4**	**1**	**2**
$X = 4$	**2**	**3**	**4**	**1**

Etwa bei $X = 2$ und $W = 1$ ist die Antwort $A = 4$ zu geben. Aus der Antwort A lässt sich der Wert von X nicht identifizieren. Es sind sogar für jede Antwort A noch alle X-Werte

möglich. Da jeder Befragte lediglich eine verschlüsselte Antwort A zu geben hat und nicht seinen Wert von X preisgeben muss, ist die Privatsphäre geschützt. Folglich ist davon auszugehen, dass die Kooperationsbereitschaft bei einer Umfrage mit dem DM höher ist als bei Direktbefragung.

Das eben beschriebene DM ist ein "Nonrandomized-Response"-Umfrageverfahren (kurz NRR-Verfahren). Das bedeutet, wenn eine Person mehrfach befragt wird, so erhält man stets dieselbe Antwort A. Im Gegensatz dazu sind in der Literatur auch "Randomized-Response"-Methoden (RR-Methoden) bekannt. Bei diesen hängt die zu gebende Antwort eines Interviewten neben dessen Wert von X auch vom Ergebnis eines Zufallsexperimentes ab. Wird also bei einem RR-Design eine Person mehrfach in die Stichprobe gezogen, so sind unterschiedliche Antworten möglich.

Die Entwicklung des DM war motiviert durch einige Nachteile von zuvor zwischen 2007 und 2009 in hochrangigen Journals publizierten NRR-Techniken. Im Artikel Groenitz (2012) wird zunächst auf die Limitierungen von anderen NRR-Verfahren eingegangen und anschließend der Ablauf einer Umfrage gemäß DM dargestellt.

Anschließend wird darauf eingegangen, wie man aus den beobachteten Antworten gemäß DM Rückschlüsse auf die Verteilung von X zieht. Dabei gehen wir davon aus, dass eine Stichprobe gemäß einfacher Zufallsauswahl mit Zurücklegen (simple random sampling with replacement, SRSWR) vorliegt. Einfache Zufallsauswahl bedeutet, dass jede mögliche Stichprobe die gleiche Auswahlwahrscheinlichkeit hat. Offenbar lässt sich die Verteilung von X durch einen Vektor π der Länge k beschreiben, wobei die i-te Komponente von π den Anteil der Personen in der Population mit Ausprägung $X = i$ repräsentiert. Analog lässt sich die Verteilung von W bzw. A durch einen Vektor $c = (c_1, ..., c_k)$ bzw. $\lambda = (\lambda_1, ..., \lambda_k)^T$ beschreiben. Hierbei ist c_i bzw. λ_i der Anteil der Personen in der Grundgesamtheit, die den Merkmalswert $W = i$ bzw. $A = i$ besitzen.

Es wird die Maximum-Likelihood-Schätzung (ML-Schätzung) für π beschrieben und gezeigt, dass der EM-Algorithmus nutzbringend zur Berechnung von ML-Schätzwerten ist. Der EM-Algorithmus ist eine in der Literatur bekannte Methode zur Berechnung von ML-Schätzern in Missing-Data-Problemen, d.h. bei Datensätzen mit fehlenden Werten. Die entscheidende Beobachtung, welche die Anwendbarkeit des EM-Algorithmus in unserer Situation sicherstellt, ist, dass eine Umfrage gemäß DM auf eine spezielle Missing-Data-Konstellation führt: Die Werte von X sind nie beobachtet (diese Werte sind die fehlenden Werte), wohingegen die Realisierungen von A die beobachteten Werte darstellen. Mit dem EM-Algorithmus sind wir stets in der Lage einen zulässigen Schätzer $\hat{\pi}$ für π (d.h. alle Komponenten des Schätzers sind zwischen 0 und 1, die Summe der Komponenten ist gleich 1) anzugeben. In diesem Zusammenhang halten wir fest, dass in vielen Publikationen anderer Autoren zu RR/NRR-Designs das Problem von unzulässigen Schätzern nicht zufriedenstellend gelöst wird oder gar nicht auf das Problem eingegangen wird.

Im Abschnitt 3.3 in Groenitz (2012) werden die geschätzten Standardfehler der Schätzung angegeben sowie asymptotische und Bootstrap-Konfidenzintervalle hergeleitet und verglichen.

Danach folgt eine ausführliche Diskussion von Effizienz der Schätzung und dem Grad

an Schutz der Privatsphäre (degree of privacy protection, DPP). Hohe Effizienz bedeutet geringe Schätzungenauigkeit. Die Schätzungenauigkeit messen wir mit der Summe der MSEs der Komponenten von $\hat{\pi}$ (MSE: mean squared error, also mittlerer quadratischer Schätzfehler). Es zeigt sich, dass sich die Schätzungenauigkeit für das DM zusammensetzt aus der Schätzungenauigkeit, die man bei Direktbefragung und wahren Antworten ohne Antwortverweigerungen hätte, plus einem Aufschlag für die indirekte Befragung gemäß DM. Die Schätzungenauigkeit bei Direktbefragung hängt hierbei von π ab, der Aufschlag ist abhängig von c. Dieser Aufschlag kann interpretiert werden als Preis, der für den Schutz der Privatsphäre der Befragten gezahlt wird.

Wir kommen nun zur Messung des DPP. Wenn W eine Einpunktverteilung hätte (d.h. eine Komponente von c ist gleich 1, die anderen Komponenten sind alle gleich 0), wäre die Privatsphäre überhaupt nicht geschützt, denn man könnte aus A den Wert von X rekonstruieren. Andererseits, der größtmögliche Schutz der Privatsphäre der Befragten liegt vor, falls W eine Gleichverteilung besitzt (also alle Einträge von c gleich $1/k$ sind). In diesem Fall sind A und X unabhängig. Um den DPP zu messen, bietet es sich gemäß der eben skizzierten Überlegungen an, zu betrachten, wie weit die Verteilung von W von einer Gleichverteilung und einer Einpunktverteilung entfernt ist. Daher quantifizieren wir den DPP über die Standardabweichung σ des Vektors c. Ist σ groß, so ist die Verteilung von W nahe einer Einpunktverteilung (also der DPP klein) während ein kleiner Wert von σ anzeigt, dass die Verteilung von W nahe an einer Gleichverteilung liegt und somit ein großer DPP verfügbar ist.

In der Arbeit Groenitz (2012) wird gezeigt, dass der Aufschlag bei der Schätzungenauigkeit für das DM eine DPP-abhängige Untergrenze besitzt. Das bedeutet, es gibt optimale und nicht-optimale Vektoren c. Ein c ist nicht optimal, falls es einen gewissen DPP σ liefert, aber zu einem Aufschlag der Schätzungenauigkeit führt, der größer ist als für dieses σ notwendig. Es wird weiterhin hergeleitet, wie man zu einem optimalen Vektor c für einen vorgegebenen DPP kommt. Wenn man schließlich nur optimale Vektoren c betrachtet, so ist der Aufschlag bei der Schätzungenauigkeit eine streng monoton fallende Funktion von σ. Das bedeutet, je mehr Schutz der Privatsphäre den Interviewten gegeben wird, desto höher ist der Aufschlag bei der Schätzungenauigkeit. Folglich muss eine Abwägung getroffen werden: Ein gewisse Maß an Schutz der Privatsphäre muss den Befragten zugestanden werden, um deren Kooperation zu sichern, bei zu viel Schutz jedoch leidet die Präzision der Schätzung. In der Praxis ist es daher sinnvoll, ein mittleres σ auszuwählen, hierzu einen optimalen Vektor c festzulegen und schließlich ein Merkmal W an dieses c anzupassen.

Es sei hier ausdrücklich darauf hingewiesen, dass Resultate über den Zusammenhang DPP / Effizienz wie in Groenitz (2012) (mathematische Funktion für die Abhängigkeit des Aufschlages bei der Schätzungenauigkeit vom DPP, Herleitung von optimalen Modellparametern für jeden DPP) nur sehr selten in der Literatur über RR/NRR-Verfahren für kategoriale X (mit beliebig vielen Kategorien) zu finden sind.

Die Manuskripte Groenitz (2013a), Groenitz (2013b) und Groenitz (2013c) stellen Erweiterungen zur Arbeit von Groenitz (2012) vor.

Im Essay Groenitz (2013a) wird wieder ein kategoriales, sensitives Merkmal X betrachtet und angenommen, dass Daten über X mit Hilfe des DM gesammelt wurden (d.h. es liegen

verschlüsselte Antworten A vor). Dabei gehen wir wieder von einer Stichprobe gezogen durch SRSWR aus. Es wird nun der Fall untersucht, bei dem Vorinformation über die Verteilung von X verfügbar ist. Die Vorinformation könnte z.B. aus einer vorangegangenen Studie stammen. Um die Vorinformation in die Schätzung der Verteilung von X einzubeziehen, bieten sich Bayesianische Methoden an. Bei Bayesianischen Schätzverfahren wird die Vorinformation in einer "priori"-Verteilung gesammelt und die "posteriori"-Verteilung analysiert. Die in der posteriori-Verteilung enthaltene Information setzt sich zusammen aus der Vorinformation und der Information aus den erfassten Antworten der aktuellen Umfrage.

Es gibt verschiedene Möglichkeiten, die posteriori-Verteilung auszuwerten, jede davon liefert einen etwas anderen Schätzer für die Verteilung von X. Im Einzelnen werden im Artikel Groenitz (2013a) der Modus der posteriori-Verteilung des Parameters sowie Schätzer basierend auf Parameter-Simulation, multipler Imputation und Rao-Blackwellisierung ermittelt. Für die drei letztgenannten Methoden ist der Data-Augmentation-Algorithmus, welcher gewisse Markov-Ketten generiert, hilfreich. Ein Vergleich der betrachteten Bayes-Schätzverfahren beschließt den ersten Teil des Manuskriptes von Groenitz (2013a).

Bei der Berechnung von Bayes-Schätzern für das DM fällt auf, dass die Designmatrix des DM (dies ist eine Matrix, deren Einträge gewisse Wahrscheinlichkeiten sind) hier die zentrale Rolle spielt. Im zweiten Teil des Aufsatzes Groenitz (2013a) wird die folgende Verallgemeinerung dieser Beobachtung bewiesen: Für jedes RR- oder NRR-Modell, das kategoriale Merkmale behandelt, ist die Menge der Designmatrizen des Modells die einzige Komponente des Modells, die für die Bayes-Schätzung gebraucht wird. Das konkrete Antwortschema wird nicht benötigt. Dieses Resultat ermöglicht die umfangreiche Verallgemeinerung der Formeln aus dem ersten Teil und die Etablierung eines gemeinsamen Ansatzes für die Bayes-Schätzung in RR-/ NRR-Modellen für kategoriale Merkmale. Dieser vereinheitlichte Ansatz deckt viele vorhandene und potentielle RR-/ NRR-Designs einschließlich gewisse mehrstufige Designs und Designs, die mehrere Stichproben benötigen, ab.

Wie oben beschrieben, präsentiert der Artikel Groenitz (2012) die Schätzung der Verteilung eines sensitiven, kategorialen Merkmals X basierend auf den DM-Antworten von sagen wir n Personen. In diesem Artikel wird dabei unterstellt, dass die n Befragten durch einfache Zufallsauswahl mit Zurücklegen ausgewählt wurden. In der Praxis werden jedoch auch andere Stichprobenziehungen als SRSWR verwendet. Dies motiviert den Aufsatz Groenitz (2013b), in welchem Schätzer für das DM für weitere wichtige Stichprobenziehungen entwickelt werden. Dabei wird auf geschichtete Stichproben, Stichproben mit unterschiedlichen Auswahlwahrscheinlichkeiten, Klumpen-Stichproben und mehrstufige Stichproben jeweils für Ziehen mit als auch ohne Zurücklegen eingegangen. Für jedes betrachtete Stichprobenauswahlverfahren untersuchen wir auch die Eigenschaften des hergeleiteten Schätzers wie Varianz und den Zusammenhang zwischen Grad an Schutz der Privatsphäre und Effizienz.

Das Manuskript Groenitz (2013c) betrachtet eine Umfrage mit einem sensitiven, kategorialen Merkmals Y^*, das die möglichen Werte $1, ..., k$ besitzt, und nicht-sensitiven Kovariablen $X_1^*, ..., X_p^*$. Beachte, um der Notation in Groenitz (2013c) zu folgen, bezeichnen wir das sensitive Merkmal ab hier mit Y^*. Es wird davon ausgegangen, dass die Daten über Y^* mit Hilfe des DM aus Groenitz (2012) gesammelt werden. Das Ziel ist es nun,

Methoden zu entwickeln, mit denen man den Einfluss von $X^* = (X_1^*, ..., X_p^*)$ auf Y^* untersuchen kann. Zum Beispiel, wenn Y^* Einkommensklassen repräsentiert, könnte man sich für die Abhängigkeit des Merkmals Y^* von den Kovariablen Geschlecht (X_1^*) und Beruf (X_2^*) interessieren. Im Aufsatz Groenitz (2013c) werden sowohl deterministische als auch stochastische Kovariablen behandelt. Legt der Forscher die Werte von X^* fest und sucht dann Personen, die die ausgewählten Kovariablenlevel besitzen, liegen deterministische Kovariablen vor. In diesem Fall wird jede ausgewählte Person gebeten, eine Antwort A^* gemäß dem Diagonal-Modell zu geben, d.h. A^* hängt von Y^* und einem Hilfsmerkmal W^* ab. Andererseits, sobald man Personen in die Stichprobe auswählt, ohne vorher Werte von X^* festzulegen, haben wir stochastische Kovariablen, also zufällige Werte von X^*. Im Falle stochastischer Kovariablen werden bei jedem Interview zuerst die Werte von $X_1^*, ..., X_p^*$ direkt erfragt (sofern diese nicht bereits offensichtlich sind wie z.B. beim Geschlecht). Anschließend wird um eine Antwort gemäß DM gebeten.

Im Artikel Groenitz (2013c), Abschnitt 3.1, werden deterministische Kovariablen betrachtet. Hierbei wird zunächst die schichtweise Schätzung beschrieben. Diese ist geeignet, wenn hinreichend viele Beobachtungen für jedes der aufgetretenen Kovariablenlevel vorliegen. Der Schwerpunkt der Arbeit liegt allerdings auf der Herleitung von "LR-DM-Schätzern" und der Untersuchung von Eigenschaften dieser Schätzer. Dabei ist ein "LR-DM-Schätzer" ein Schätzer, der auf der Annahme eines logistischen Regressionsmodells für die Beziehung zwischen Y^* und X^* basiert. Bei der LR-DM-Schätzung werden vielfältige Methoden aus dem Bereich der Generalisierten Linearen Modelle benötigt (z.B. der Fisher-Scoring-Algorithmus zur iterativen Berechnung des Schätzers).

Im anschließenden Abschnitt 3.2 wird erläutert, wie die Methoden und Erkenntnisse für deterministische Kovariablen auf den Fall stochastischer Kovariablen übertragen werden können. Zum Aufsatz Groenitz (2013c) gehört auch ein Abschnitt mit umfangreichen Simulationen. In diesen wird die Beziehung zwischen Grad an Schutz der Privatsphäre und Effizienz des LR-DM-Schätzers analysiert sowie die Präzision von LR-DM-Schätzung und schichtweiser Schätzung verglichen.

Die vier Artikel, auf die sich diese Zusammenfassung bezieht, involvieren zum Teil computer-intensive Methoden. Aus diesem Grund sind folgende selbst-erstellten MATLAB-Programme, welche die entsprechenden Rechnungen ausführen, als Zusatzmaterial beigefügt.

- `estimationDM.m`

  Dieses Programm ist Zusatzmaterial zu Groenitz (2012). Es berechnet ML-Schätzer (ggf. über EM-Algorithmus) und gibt Konfidenzintervalle aus.

- `Bayes_est.m`

  Dieses Programm ist Beilage zu Groenitz (2013a) und ermöglicht die Ermittlung von Bayes-Schätzern für diverse RR-/ NRR-Modelle.

- `fisherscore1.m`

  Dieses Programm ist Beilage zu Groenitz (2013c) und berechnet LR-DM-Schätzer über den Fisher-Scoring-Algorithmus.

A New Privacy-Protecting Survey Design for Multichotomous Sensitive Variables.

Heiko Groenitz

Groenitz, H. (2012): A New Privacy-Protecting Survey Design for Multichotomous Sensitive Variables. Metrika, DOI: 10.1007/s00184-012-0406-8.

```
function [pi_hat, Iter, SEpsi,BT1,BT2,AS] = estimationDM(h,n,c, f,Gf,B, alpha)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%↙
%%%%%%%%%%%%%%%%

% Supplemental material for the paper
% Groenitz, H. (2012): A New Privacy-Protecting Survey Design for
% Multichotomous Sensitive Variables.
% Metrika, DOI: 10.1007/s00184-012-0406-8.

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%↙
%%%%%%%%%%%%%%%%

%DESCRIPTION:
%The function 'estimationDM' enables the estimation in the diagonal model.

%Eiter 3 or 7 input arguments are required:
%[pi_hat, Iter] = estimationDM(h,n,c) calculates the MLE pi_hat for the
%true parameter pi and returns the number of iterations in EM algorithm

%[pi_hat, Iter, SEpsi,BT1,BT2,AS] = estimationDM(h,n,c, f,Gf,B, alpha)
%additionally returns the bootstrap standard error, bootstrap confidence
%intervals (CI) and an asymptotic CI for a function psi=f(pi)

%INPUT:
%h:observed relative frequencies of the answers A=1,...,A=k (column vector)
%n: sample size
%c: vector describing the distribution of the auxiliary variable W
%f: real-valued function (psi = f(pi) is a function of the true parameter)
%Gf: gradient of f; Gf: R^k --> R^k;
%B: Number of bootstrap replications
%1-alpha: confidence level

%OUTPUT:
%pi_hat: calculated estimator for pi
%Iter: number of iterations of EM algorithm
%(if Iter=0, EM algorithm was not necessary)
%SEpsi: estimated standard error for psi (with bootstrap)
%BT1 / BT2: bootstrap CI's (with / without normality assumption)
%AS: asymptotic confidence interval (CI) for psi (via delta method)


%EXAMPLE:
%Let the following frequencies of the answers A=1,...,A=4 be
%observed: (n_1,...,n_4)=[63 45 73 69]'.
%
% nn=[63 45 73 69]'; n=sum(nn);h=nn/n; c=[0.625 0.125 0.125 0.125]
% f=@(x)x(1); Gf=@(x)[1;0;0;0]; B=2000; alpha=0.05
%
% r e s u l t s:
% pi_hat = [0.2540  0.3020  0.3340  0.1100]', Iter = 0,
% SEpsi = 0.0551, BT1 = [0.1460  0.3620], BT2 = [0.1500  0.3660],
```

```
% AS = [0.1464  0.3616]

%-------------------------------------------------------------------
% nested function (for calculation of pi_hat)

function [pi_hat,Iter]=pi_hatEM_DM(h,n,C_0,k)

% Calculate inv(C_0)*h

pi_hat=C_0\h; % [= inv(C_0)*h]

if (pi_hat>=0) & (pi_hat<=1)% Check if 0<= pi_i <= 1
T=1; Iter=0;
else
   T=2; % EM algorithm necessary
end


if T==2   %run EM algorithm

  p1= ones(k,1)/k; % initial value
   %E step

   A=C_0*p1;
   M=( C_0*( (n*h)./A) ).*p1;

   %M step: new estimator
   p2=M/sum(M);     Iter=1;

   while max(abs(p2-p1)) > 10^-8
      Iter=Iter+1;
      p1=p2;

      %E step

      A=C_0*p1;
      M=( C_0*( (n*h)./A) ).*p1;

      %M step: new estimator
      p2=M/sum(M);

   end

   pi_hat=p2;
end

end

%-------------------------------------------------------------------

k=length(c);
```

```
% Calculation of the design matrix C_0 induced by c

CIR=gallery('circul',c); %CIR is a circulant matrix
C_0(1,:)=CIR(1,:); C_0(2:k,:)=flipud(CIR(2:k,:));

%-------------------------------------------------------------------

% Computation of the estimator pi_hat

[pi_hat,Iter]=pi_hatEM_DM(h,n,C_0,k);


%-------------------------------------------------------------------
if nargin==3

    SEpsi='NA';
    BT1='NA';
    BT2='NA';
    AS='NA';

elseif nargin==7    %calculate SEpsi,BT1,BT2,AS

la_hat=C_0*pi_hat; %estimated answer probabilities
psi_hat=feval( f, pi_hat);

% Bootstrap standard error and bootstrap confidence intervals for psi

PSI=zeros(B,1); %collects bootstrap replications of psi_hat
for i=1:B
    nn=mnrnd(n,la_hat)';  %new answer frequencies
    [p,It]=pi_hatEM_DM(nn/n,n,C_0,k); %new MLE p
    PSI(i)=feval(f,p); %i-th replication psi^(i)
end

SEpsi=std(PSI);    %bootstrap standard error

% Bootstrap CI for psi with normality assumption

q=norminv(1-alpha/2);
BT1=[psi_hat-q*SEpsi      psi_hat+q*SEpsi];

% Bootstrap CI for psi without normality assumption

BT2=[quantile(PSI,alpha/2)    quantile(PSI,1-alpha/2)];


% Asymptotic CI (delta method) for psi

GA_hat=inv(C_0)*diag(la_hat)*inv(C_0) - diag(pi_hat);    %Gamma
DE_hat=diag(pi_hat) - pi_hat*pi_hat';                    %Delta
V_hat=1/n * (GA_hat+DE_hat);
```

```
Spsi=sqrt( feval(Gf,pi_hat)' * V_hat * feval(Gf,pi_hat) );

AS=[psi_hat-q*Spsi      psi_hat+q*Spsi];

else error('Number of input arguments must be 3 or 7')
end

end
```

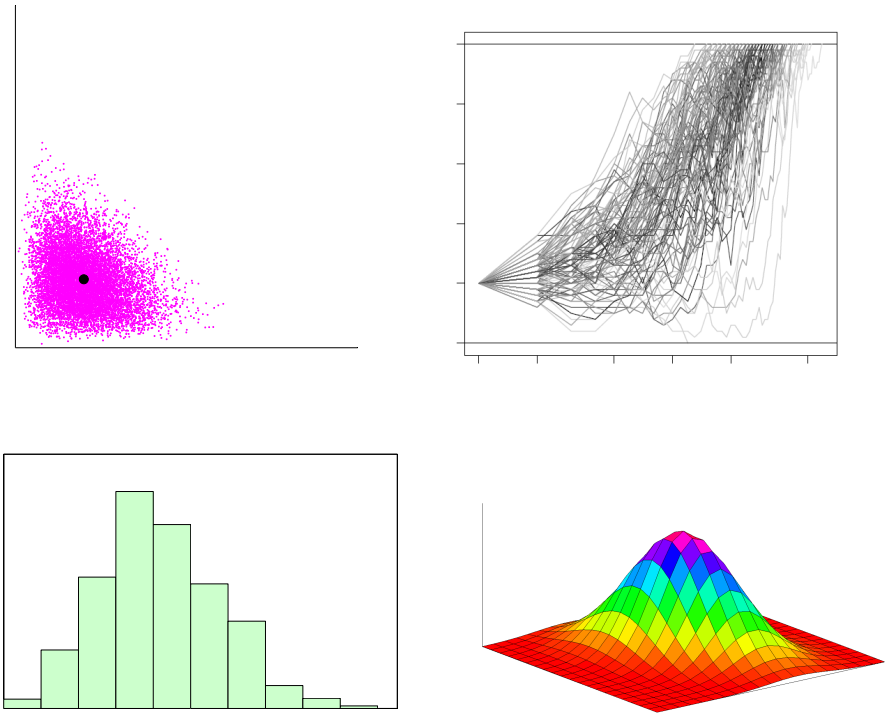
Philipps Universität Marburg

Discussion Papers on Statistics and Quantitative Methods

Using Prior Information in Privacy-Protecting Survey Designs for Categorical Sensitive Variables

Heiko Groenitz

1 / 2013



Download from:
http://www.uni-marburg.de/fb02/statistik/forschung/discpap

Coordination: Prof. Dr. Karlheinz Fleischer • Philipps-University Marburg
Faculty of Business Administration • Department of Statistics
Universitätsstraße 25 • D-35037 Marburg
E-Mail: k.fleischer@wiwi.uni-marburg.de

Using Prior Information in Privacy-Protecting Survey Designs for Categorical Sensitive Variables

Heiko Groenitz[1]

02.01.2013

Abstract

To gather data on sensitive characteristics, such as annual income, tax evasion, insurance fraud or students' cheating behavior, direct questioning is not helpful, because it results in answer refusal or untruthful responses. For this reason, several randomized response (RR) and nonrandomized response (NRR) survey designs, which increase cooperation by protecting the respondents' privacy, have been proposed in the literature. In the first part of this paper, we present a Bayesian extension of a recently published, innovative NRR method for multichotomous sensitive variables. With this extension, the investigator is able to incorporate prior information on the parameter, e.g. based on a previous study, into the estimation and to improve the estimation precision. In particular, we calculate posterior modes with the EM algorithm as well as estimates based on parameter simulation, multiple imputation, and Rao-Blackwellization. The performance of these estimation methods is evaluated in a simulation study. In the second part of this article, we show that for any RR or NRR model, the design matrices of the model play the central role for the Bayes estimation whereas the concrete answer scheme is irrelevant. This observation enables us to widely generalize the calculations from the first part and to establish a common approach for the Bayes inference in RR and NRR designs for categorical sensitive variables. This unified approach covers even multi-stage models and models that require more than one sample.

Zusammenfassung

Zur Datenerhebung bei sensitiven Merkmalen wie Einkommen, Steuerhinterziehung, Versicherungsbetrug oder Prüfungsbetrug ist Direktbefragung problematisch, da sie oft zu Antwortverweigerungen oder Falschantworten führt. Aus diesem Grund wurden in der Literatur verschiedene Randomized-Response- und Nonrandomized-Response-Umfrageverfahren (kurz RR- und NRR-Verfahren), welche die Privatsphäre der Befragten schützen und dadurch deren Kooperationsbereitschaft erhöhen, vorgeschlagen. Im ersten Teil dieses Aufsatzes präsentieren wir eine Bayes-Erweiterung eines kürzlich publizierten NRR-Modells für kategoriale sensitive Merkmale. Durch diese Erweiterung ist es möglich Vorinformation über den Parameter, die zum Beispiel auf einer vorherigen Erhebung basieren könnte, in die Schätzung einzubeziehen und dadurch die Schätzgenauigkeit zu verbessern. Wir ermitteln den Modus der a-posteriori-Verteilung mit dem EM-Algorithmus und berechnen Schätzer basierend auf Parametersimulation, multipler Imputation und Rao-Blackwellisierung. Diese Schätzverfahren werden im Rahmen einer Simulationsstudie verglichen. Im zweiten Teil des Artikels zeigen wir, dass die Designmatrizen des Modells bei jedem RR- / NRR-Modell für kategoriale sensitive Merkmale die zentrale Rolle für die Bayes-Schätzung spielen wohingegen die konkrete Antwortformel irrelevant ist. Diese Beobachtung ermöglicht es uns die Rechnungen aus dem ersten Teil des Aufsatzes weitreichend zu verallgemeinern und einen gemeinsamen Ansatz für die Bayes-Schätzung bei RR- / NRR-Verfahren zu entwickeln. Dieser vereinheitlichte Ansatz deckt sogar mehrstufige Modelle sowie Modelle, welche mehrere Stichproben benötigen, ab.

KEYWORDS: Randomized response; Nonrandomized response; Bayesian estimation; EM algorithm; Data augmentation

[1]Philipps-University Marburg, Department for Statistics (Faculty 02), Universitätsstraße 25, 35032 Marburg, Germany (e-mail: groenitz@staff.uni-marburg.de).

1 Introduction

Let us consider a survey on a sensitive attribute X. For instance, X may represent income classes or the number of times the respondent has evaded taxes. In the case of direct questioning (DQ), many respondents will not reveal the true value of X. Instead, answer refusal and untruthful responses will occur. This leads to a serious bias when estimating the distribution of X based on DQ. For this reason, several randomized response (RR) and nonrandomized response (NRR) techniques have been developed in the literature to obtain trustworthy estimates of the distribution of X. To protect privacy, the respondents are always requested to provide a scrambled answer A instead of the X-value. This practice reduces untruthful answers and answer refusal. The realizations of A and X are observed and missing data, respectively.

A RR technique was first proposed by Warner (1965), whose seminal model has been extended in various dimensions until today. RR models have in common that every respondent is supplied with a randomization device (RD), such as a coin or a deck of cards. The respondents use the RD to conduct a random experiment, whose outcome influences the required scrambled answer. The necessity of running the random experiment is cumbersome. This is why nonrandomized response approaches are coming up in recent years with articles by Tian et al. (2007), Yu et al. (2008), Tan et al. (2009), Tang et al (2009) and Groenitz (2012). NRR models do not need a RD; in such models, the answer depends on an auxiliary variable, and the respondent would give the same answer if he or she was asked again. NRR methods are easy to implement and suitable for face-to-face and e-mail surveys. Compared with RR techniques, NRR methods reduce both survey complexity and study costs.

In privacy-protecting (PP) models (i.e., RR or NRR designs), maximum likelihood (ML) estimates can be derived from the empirical distribution of the scrambled answers. However, for the case in which prior information on the distribution of interest is available, Bayesian methods should be applied to incorporate the prior information. Bayesian estimation means that we collect the prior information in a prior distribution and analyze the observed data posterior distribution. Note that even if there is no prior information, the Bayesian approach with a uniform prior distribution can be recommendable: for this prior, the posterior mode equals the ML estimator (MLE). However, in small samples, the posterior standard deviation and confidence intervals based on posterior quantiles can be expected to be more suitable than the asymptotic standard error of the MLE and confidence intervals based on the asymptotic normality of the MLE.

Bayesian methods (usually based on a Dirichlet prior) have been proposed for some PP designs: Winkler and Franklin (1979) as well as Migon and Tachibana (1997) present Bayesian approaches for Warner's (1965) RR model. O'Hagan (1987) derives Bayes linear estimators for Warner's model and the unrelated question model (UQM) by Horvitz et al. (1967). Unnikrishnan and Kunte (1999) describe a unified model for Warner's model and the UQM as well as a unified model for the common handling of the model by Abul-Ela et al. (1967) and the polychotomous UQM by Greenberg et al. (1969). For both unified models, the Gibbs sampler is used to generate realizations from the posterior distribution. Bayesian inference for Mangat's (1994) RR model can be found in Kim et al. (2006). Tang et al. (2009) suggest a certain NRR model and explain the corresponding Bayesian estimation. Bayesian methods for the NRR methods by Tian et al. (2007) and Yu et al. (2008) can be found in Tian et al. (2009). Barabesi and Marcheselli (2010) propose a Bayesian approach to the joint estimation of the distribution of a binary sensitive variable and the sensitivity level from data collected with a certain two-stage RR scheme. The Bayes estimation for the RR model by Mangat and Singh (1990) is derived in Hussain et al. (2011).

In the first part of this paper, we extend the work by Groenitz (2012), who presents the nonrandomized diagonal model (DM) including ML estimation, in order to have the possibility to incorporate prior information into the estimation and to obtain more precise estimates. In Section 2, we narrate

the diagonal model and derive Bayesian estimates for this model. In particular, we calculate posterior modes via the EM algorithm as well as estimates based on parameter simulation (PS), multiple imputation (MI) and Rao-Blackwellization (RB) for the DM survey design. For PS, MI, RB, the data augmentation algorithm, which generates certain Markov chains, turns out to be beneficial. The quality of PS, MI, RB for a survey according to the diagonal model is investigated in a simulation study.

For the DM, we observe in Section 2 that the design matrix of the model, i.e., a matrix of conditional probabilities, plays the central role for the calculation of posterior modes and estimates based on PS, MI, RB. In the second part of this paper, we show the following generalization of this observation: For any PP survey model dealing with categorical X, the only component of the model that is needed to compute Bayes estimates is the set of design matrices of the model. The concrete answer scheme is irrelevant for Bayes inference. This result enables us to establish a common approach for the Bayes estimation in PP survey designs for categorical sensitive variables in Section 3. This unified approach covers many published and potential PP designs including certain multi-stage designs and designs demanding multiple samples. Here, we derive general formulas that can be applied to a lot of PP models for which Bayesian concepts have not been discussed yet.

2 Bayes estimation for the diagonal model

2.1 Diagonal model

Groenitz (2012) proposed the diagonal model (DM), which can be applied to gather data on a sensitive characteristic $X \in \{1, ..., k\}$. For the DM, a nonsensitive auxiliary variable $W \in \{1, ..., k\}$ (e.g., W may describe the period of birthday) must be specified such that X and W are independent and that the distribution of W is known. The respondent is introduced to give the answer

$$A := [(W - X) \bmod k] + 1. \tag{1}$$

Equation (1) should not be shown to the respondents; instead, every interviewee receives a table that illustrates (1). E.g., for $k = 4$, we have

X/W	$W = 1$	$W = 2$	$W = 3$	$W = 4$
$X = 1$	**1**	**2**	**3**	**4**
$X = 2$	**4**	**1**	**2**	**3**
$X = 3$	**3**	**4**	**1**	**2**
$X = 4$	**2**	**3**	**4**	**1**

The number in the interior of the table is the required answer A. Notice, the answers A do not restrict the possible X-values. Hence, we assume that the interviewees cooperate and reveal their values of A. We remark that the DM is applicable even if all the values of X are sensitive (e.g., if the values of X correspond to income classes).

Throughout this article, let π_i, c_i, λ_i be the proportion of units in the population having attribute $X = i$, $W = i$, $A = i$, respectively. Moreover, define $C(i, j)$ to be the proportion of individuals having $A = i$ among the persons with $X = j$. We then have $(\lambda_1, ..., \lambda_k)^T = C \cdot (\pi_1, ..., \pi_k)^T$ with the $k \times k$ matrix $C = [C(i, j)]_{ij}$, where every row of C is a left-cyclic shift of the row above and the first row of C is equal to $(c_1, ..., c_k)$. C is called the "design matrix" and plays an important role for the Bayes estimation in the DM.

2.2 Basic principles and definitions for Bayes estimation

We assume a simple random sample with replacement (SRSWR) of n units has been drawn. These n persons are introduced to answer according to the DM answer formula (1). Let X_i and A_i be the i-th respondent's value of X and A, respectively. Consequently, $\mathbf{A} = (A_1, ..., A_n)$ and $\mathbf{X} = (X_1, ..., X_n)$ represent the observed data and the missing data, respectively. Thus, a DM survey generates a data structure that corresponds to a special missing data problem. For this reason, we can apply known missing data methods, e.g., EM algorithm or data augmentation, to incorporate prior information into the estimation for the DM.

In the subsequent subsections, we derive Bayes estimates for the unknown $\pi = (\pi_1, ..., \pi_{k-1})^T \in \mathbb{R}^{k-1}$. In a Bayesian view, π is treated as a realization of a random variable Π. The prior information about π is collected in a prior distribution defined by a density f_Π, which is specified by the investigator. In this article, we focus on Dirichlet prior distributions. In Subsection 2.3, we explain a possibility to convert prior information into a concrete Dirichlet distribution. In addition to f_Π, the conditional distribution of the complete data $(\mathbf{X}, \mathbf{A})$ given Π must be defined. We denote the corresponding density by $f_{\mathbf{X},\mathbf{A}\,|\,\Pi}(\cdot, \cdot\,|\,\pi)$, and set for $x_j, a_j \in \{1, ..., k\}$

$$f_{\mathbf{X},\mathbf{A}\,|\,\Pi}(\mathbf{x}, \mathbf{a}\,|\,\pi) = \prod_{j=1}^{n} C(a_j, x_j) \cdot \pi_{x_j}, \tag{2}$$

where $\mathbf{x} = (x_j)_j$, $\mathbf{a} = (a_j)_j$. That is, we have conditional independence of the n vectors (X_j, A_j) given Π. It follows that

$$f_{\mathbf{X}\,|\,\mathbf{A},\Pi}(\mathbf{x}\,|\,\mathbf{a}, \pi) = \prod_{j=1}^{n} \frac{C(a_j, x_j) \cdot \pi_{x_j}}{f_{A_j\,|\,\Pi}(a_j\,|\,\pi)}, \tag{3}$$

where $f_{A_j\,|\,\Pi}(\alpha\,|\,\pi)$ is the entry number $\alpha \in \{1, ..., k\}$ of vector $C \cdot (\pi_1, ..., \pi_k)^T$.

Assume a value $\mathbf{a}$ of $\mathbf{A}$ has been observed in the survey. The basic idea is to evaluate the posterior distribution of Π given $\mathbf{a}$ and the distribution of $\mathbf{X}$ given $\mathbf{a}$. In Subsection 2.4, we compute posterior modes with the EM algorithm, and in 2.5, we describe ways based on the data augmentation algorithm (in particular, parameter simulation and multiple imputation) to estimate the true proportion π. Estimators derived by the idea of Rao-Blackwell's theorem are considered in 2.6.

2.3 Dirichlet prior distributions

The random vector $\Pi = (\Pi_1, ..., \Pi_{k-1})$ is Dirichlet distributed if it has Lebesgue density

$$f_\Pi(\pi) = f_\Pi(\pi_1, ..., \pi_{k-1}) = K \cdot \pi_1^{\delta_1 - 1} \cdots \pi_{k-1}^{\delta_{k-1} - 1} \cdot (1 - \sum_{i=1}^{k-1} \pi_i)^{\delta_k - 1} \cdot 1_{E_{k-1}}(\pi), \tag{4}$$

where $E_{k-1} = \{(x_1, ..., x_{k-1}) \in [0,1]^{k-1} \,:\, x_1 + ... + x_{k-1} \leq 1\}$, $\delta = (\delta_1, ..., \delta_k)$ is a vector of parameters with $\delta_i > 0$ and K is a constant depending on δ. We will usually write $\Pi \sim Di(\delta)$ in the sequel. Let us assume that $(\hat{\pi}_1^{(p)}, ..., \hat{\pi}_k^{(p)})^T$ is the investigator's guess for the unknown proportions. This guess may be based on a previous study. One option to convert this guess into a Dirichlet distribution is as follows. Choose a proportionality factor d, and define δ_i to be proportional to $\hat{\pi}_i^{(p)}$, i.e, $\delta_i = \hat{\pi}_i^{(p)} \cdot d$. Let $(D_1, ..., D_{k-1})$ be Dirichlet distributed with these δ_i. Then, we have $\mathbb{E}(D_i) = \hat{\pi}_i^{(p)}$ and $Var(D_i) = \hat{\pi}_i^{(p)}(1 - \hat{\pi}_i^{(p)})/(d+1)$. Obviously, small and large d result in a large and small variance, respectively. If the investigator feels certain that his or her guess is close to the true vector of proportions for the current study, a relatively large d should be chosen. If the investigator is unsure, a relatively small d will reflect this uncertainty.

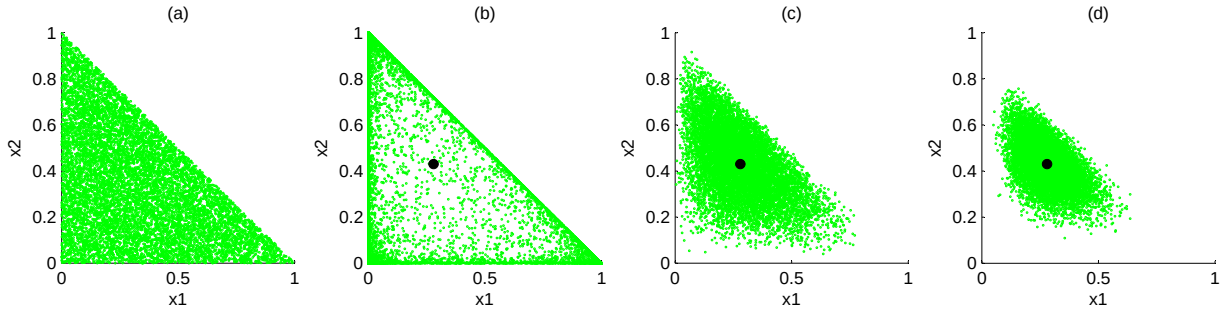


Figure 1: *Scatter plots of each 10000 random numbers from several Dirichlet distributions. In (a), we have* $\delta = (1, 1, 1)$*, for (b)-(c) we use* δ_i *as described in Subsection 2.3 where* $d = 0.5$ *in (b),* $d = 10$ *in (c) and* $d = 25$ *in (d). The black point equals* (0.28, 0.43)*, which is the investigator's guess for the unknown* π_1 *and* π_2*.*

The scatter plots of each 10000 draws from several Dirichlet distributions for $k = 3$ can be found in Figure 1. Realizations of the Dirichlet distribution can be obtained from Gamma distributed random variables, see Gentle (1998), p. 111. For $\delta = (1, 1, 1)$, the points (x_1, x_2) are uniformly scattered on E_2. This corresponds to a situation without prior information. For the figures (b) - (d), we define (0.28, 0.43, 0.29) to be the investigator's guess. In (b), we use $d = 0.5$ and δ_i as described above. It seems that there are more realizations close to the boundaries $x_1 = 0$, $x_2 = 0$, and $x_1 + x_2 = 1$ than realizations close to (0.28, 0.43). Thus, $d = 0.5$ seems inappropriate. In (c), we have $d = 10$, and the draws form a point cloud around (0.28, 0.43). The extent of this point cloud is larger than the extent of the point cloud in (d) where $d = 25$. That is, situation (d) corresponds to a larger certainty concerning the guess for the unknown true proportions.

2.4 Posterior modes for the diagonal model

As described in Dempster, Laird, Rubin (1977) for general missing data situations, the EM algorithm can be applied to generate a sequence $\pi^{(t)}$ that converges to the posterior mode, i.e, the mode of the observed data posterior density $f_{\Pi \mid \mathbf{A}}(\cdot \mid \mathbf{a})$. In particular, we have

$$\log f_{\Pi \mid \mathbf{X},\mathbf{A}}(\pi \mid \mathbf{x}, \mathbf{a}) = \log f_{\mathbf{A} \mid \Pi}(\mathbf{a} \mid \pi) + \log f_{\mathbf{X} \mid \mathbf{A},\Pi}(\mathbf{x} \mid \mathbf{a}, \pi) + \log f_{\Pi}(\pi) + \textit{constant}. \tag{5}$$

Let $\pi^{(t)}$ be available from iteration t. Computing the expectation with respect to the distribution given by $f_{\mathbf{X} \mid \mathbf{A},\Pi}(\cdot \mid \mathbf{a}, \pi^{(t)})$ yields

$$Q(\pi \mid \pi^{(t)}) + \log f_{\Pi}(\pi) = \log f_{\Pi \mid \mathbf{A}}(\pi \mid \mathbf{a}) + H(\pi \mid \pi^{(t)}) + \textit{constant},$$

where

$$\begin{aligned} Q(\pi \mid \pi^{(t)}) &= \int \log f_{\mathbf{X},\mathbf{A} \mid \Pi}(\mathbf{x}, \mathbf{a} \mid \pi) \cdot f_{\mathbf{X} \mid \mathbf{A},\Pi}(\mathbf{x} \mid \mathbf{a}, \pi^{(t)}) \, \partial \mathbf{x} \\ H(\pi \mid \pi^{(t)}) &= \int \log f_{\mathbf{X} \mid \mathbf{A},\Pi}(\mathbf{x} \mid \mathbf{a}, \pi) \cdot f_{\mathbf{X} \mid \mathbf{A},\Pi}(\mathbf{x} \mid \mathbf{a}, \pi^{(t)}) \, \partial \mathbf{x}. \end{aligned}$$

Notice that $Q(\pi \mid \pi^{(t)})$ equals the conditional expectation of the complete data log-likelihood given the observed data and $\pi^{(t)}$. In the E step of iteration $t + 1$, the function $Q^*(\cdot \mid \pi^{(t)})$ with $Q^*(\pi \mid \pi^{(t)}) = Q(\pi \mid \pi^{(t)}) + \log f_{\Pi}(\pi)$ is calculated. In the subsequent M step, we find $\pi^{(t+1)}$, which is the maximum of $Q^*(\cdot \mid \pi^{(t)})$. This $\pi^{(t+1)}$ increases the value of the observed data posterior density, i.e., it fulfills $f_{\Pi \mid \mathbf{A}}(\pi^{(t+1)} \mid \mathbf{a}) \geq f_{\Pi \mid \mathbf{A}}(\pi^{(t)} \mid \mathbf{a})$. A possible starting value is $(1/k, ..., 1/k)^T$. A detailed description of

this general scheme can be also found in Schafer (2000), Chapter 3.2.

Adopting this general scheme to a survey according to the diagonal model, we have for $\pi = (\pi_1, ..., \pi_{k-1})$, $\pi_k = 1 - \pi_1 - ... - \pi_{k-1}$ (apart from a constant)

$$Q(\pi \,|\, \pi^{(t)}) = \sum_{i=1}^{k} \hat{m}_i^{(t)} \cdot \log \pi_i \text{ and } Q^*(\pi \,|\, \pi^{(t)}) = \sum_{i=1}^{k} \left(\delta_i - 1 + \hat{m}_i^{(t)}\right) \cdot \log \pi_i \tag{6}$$

with $\hat{m}_i^{(t)} = \sum_{j=1}^{k} n_j \cdot \pi_i^{(t)} \cdot C(j,i) / f_{A_1 \,|\, \Pi}(j \,|\, \pi^{(t)})$, where n_j is the number of respondents in the sample giving answer j. We remark that $\hat{m}_i^{(t)}$ is equal to the sum of the i-th column of the $k \times k$ matrix

$$C \,.^* \left[\left[\tilde{n}^T ./\, \lambda(\pi^{(t)})\right] \cdot (\pi_1^{(t)}, ..., \pi_k^{(t)})\right].$$

Here, the signs $.^*$ and $./$ stand for componentwise multiplication and division, respectively, and

$$\tilde{n} = (n_1, ..., n_k) \text{ and } \lambda(\pi^{(t)}) = (f_{A_1 \,|\, \Pi}(1 \,|\, \pi^{(t)}), ..., f_{A_1 \,|\, \Pi}(k \,|\, \pi^{(t)}))^T$$

hold. The maximum of the function $Q^*(\cdot \,|\, \pi^{(t)})$ is given by $\pi_i^{(t+1)} = (\delta_i - 1 + \hat{m}_i^{(t)}) / (n - k + \delta_1 + ... + \delta_k)$.

2.5 Parameter simulation and multiple imputation for the diagonal model

Beyond finding the posterior mode, we can draw realizations from $f_{\Pi \,|\, \mathbf{A}}(\cdot \,|\, \mathbf{a})$ and $f_{\mathbf{X} \,|\, \mathbf{A}}(\cdot \,|\, \mathbf{a})$. To draw from these distributions, the data augmentation (DA) algorithm by Tanner and Wong (1987) is most convenient. The DA algorithm generates realizations $(\mathbf{x}^{(t)}, \pi^{(t)})$ of a Markov chain, short MC, $(\mathbf{X}^{(t)}, \Pi^{(t)})$ for $t \in \mathbb{N}$. This Markov chain converges in distribution to $f_{\mathbf{X}, \Pi \,|\, \mathbf{A}}(\cdot, \cdot \,|\, \mathbf{a})$. Thus, by integration, the sequence $(\Pi^{(t)})$ has the asymptotic distribution $f_{\Pi \,|\, \mathbf{A}}(\cdot \,|\, \mathbf{a})$.

Let us consider the diagonal model survey design and a prior distribution given by $f_\Pi \sim Di(\delta)$ with fixed and known parameter δ. The DA algorithm proceeds as follows. Let $\pi^{(t-1)} = (\pi_1^{(t-1)}, ..., \pi_{k-1}^{(t-1)})^T$ and $\pi_k^{(t-1)} = 1 - \sum_{i=1}^{k-1} \pi_i^{(t-1)}$ be available from the preceding iteration $t-1$. The next iteration t consists of the imputation step (I step) and the posterior step (P step):

I step: Drawing from $f_{\mathbf{X} \,|\, \mathbf{A}, \Pi}(\cdot \,|\, \mathbf{a}, \pi^{(t-1)})$ can be done by generating independent realizations x_j $(j = 1, ..., n)$, where x_j must be drawn according to the density $f_{X_j \,|\, A_j, \Pi}(\cdot \,|\, a_j, \pi^{(t-1)})$. However, we only need the frequency of value i $(i = 1, ..., k)$ among the values x_j for the subsequent P step. For this reason, let $m^{(t)}(i,j)$ describe the in iteration t simulated number of persons who have X-value j among the persons in the sample who give answer i. We draw

$$(m^{(t)}(i,1), ..., m^{(t)}(i,k)) \sim Multinomial(n_i, \gamma_i^{(t)}).$$

The vector $\gamma_i^{(t)}$ contains the cell probabilities and is defined to be the i-th row of the $k \times k$ matrix

$$C \,.^* \left[\left[(1, \cdots, 1)^T ./\, \lambda(\pi^{(t-1)})\right] \cdot \left(\pi_1^{(t-1)}, ..., \pi_k^{(t-1)}\right)\right],$$

where

$$\lambda(\pi^{(t-1)}) = (f_{A_1 \,|\, \Pi}(1 \,|\, \pi^{(t-1)}), ..., f_{A_1 \,|\, \Pi}(k \,|\, \pi^{(t-1)}))^T.$$

Set $m_j^{(t)} = \sum_{i=1}^{k} m^{(t)}(i,j)$, which is the simulated number of persons having $X = j$ in iteration t.

P step: We simulate realizations $(\pi_1^{(t)}, ..., \pi_{k-1}^{(t)})^T$ from $f_{\Pi \,|\, \mathbf{X}, \mathbf{A}}(\cdot \,|\, \mathbf{x}^{(t)}, \mathbf{a})$, which is the density corresponding to the $Di(m_1^{(t)} + \delta_1, ..., m_k^{(t)} + \delta_k)$ distribution.

To determine a starting value $\pi^{(0)}$, one option is to draw an outcome from the prior density. Alternatively, $\pi_i^{(0)} = 1/k$ can be used.

If t is large, then $\pi^{(t)}$ can be treated as realization from $f_{\Pi \mid \mathbf{A}}(\cdot \mid \mathbf{a})$. Assume we have generated one Markov chain of length $L_2 \in \mathbb{N}$. We delete $m^{(t)} = (m_1^{(t)}, ..., m_k^{(t)})$ and $\pi^{(t)}$ from the burn-in period $t = 1, ..., L_3 - 1$ and save them for $t = L_3, ..., L_2$. Thus, there remains a sequence $(m^{(t)}, \pi^{(t)})$ of length $L_2 - L_3 + 1$. We have two ways to extract information from this sequence. The first way is referred to as parameter simulation (see e.g., Schafer (2000), p. 89) and considers the $\pi^{(t)}$. The mean and the empirical standard deviation of the $\pi_i^{(t)}$ can be used as an estimate for the true proportion π_i and as a measure for the estimation precision, respectively. The empirical $\alpha/2$ and $1 - \alpha/2$ quantiles can be used as lower and upper bounds of a $1 - \alpha$ confidence interval (CI) for π_i. A slightly different strategy is to view the $m^{(t)} = (m_1^{(t)}, ..., m_k^{(t)})$, $t = L_3, ..., L_2$ as multiple imputations for the unobserved variables $(\sum_{j=1}^n 1_{\{X_j=1\}}, ..., \sum_{j=1}^n 1_{\{X_j=k\}})$. Each imputation $m^{(t)}$ results in an estimate $m^{(t)}/n$ for the unknown vector $(\pi_1, ..., \pi_k)$. That is, we obtain $L_2 - L_3 + 1$ estimates for π_i, which can be combined to a single estimate by using the mean. The empirical standard deviation and the $\alpha/2$ and $1 - \alpha/2$ quantiles of the $L_2 - L_3 + 1$ estimates for π_i are suitable to measure the estimation precision and to construct a $1 - \alpha$ CI for π_i, respectively.

In the last paragraph, we analyzed realizations of a single Markov chain, that is, we have considered a dependent sample. Of course, an alternative approach is given by simulating $L_1 \in \mathbb{N}$ independent Markov chains and saving only the values from the last iteration of each chain. It follows that we have L_1 independent draws from $f_{\Pi \mid \mathbf{A}}(\cdot \mid \mathbf{a})$ and L_1 independent multiple imputations, which can be evaluated analogously to the dependent quantities of the last paragraph.

2.6 Diagonal model estimates motivated by the Rao-Blackwell Theorem

Parameter simulation with a single Markov chain results in an estimate $s = (L_2 - L_3 + 1)^{-1} \sum_{t=L_3}^{L_2} \pi^{(t)}$ for the observed data posterior mean $\mathbb{E}(\Pi \mid \mathbf{A} = \mathbf{a})$. This s is used to estimate the true proportions π_i. In the context of a general missing data situation, Schafer (2000), section 4.2.3, discusses an estimate based on the idea of the Rao-Blackwell theorem. Applied to our situation of diagonal model interviews, this estimate is given by

$$\tilde{s} = (L_2 - L_3 + 1)^{-1} \sum_{t=L_3}^{L_2} \mathbb{E}(\Pi \mid \mathbf{X} = \mathbf{x}^{(t)}, \mathbf{A} = \mathbf{a}). \tag{7}$$

The distribution of Π given $\mathbf{a}$ and $\mathbf{x}^{(t)}$ appears in the P step of DA. Thus, we have

$$\mathbb{E}(\Pi \mid \mathbf{X} = \mathbf{x}^{(t)}, \mathbf{A} = \mathbf{a}) = \frac{(m_1^{(t)} + \delta_1, ..., m_{k-1}^{(t)} + \delta_{k-1})^T}{(n + \delta_1 + ... + \delta_k)},$$

where $m_j^{(t)}$ is again the simulated count of persons having $X = j$ in iteration t. The components of $\tilde{s}$ provide estimates for the unknown π_i. Analogously to Section 2.5, the empirical standard deviation and quantiles of $\mathbb{E}(\Pi_i \mid \mathbf{X} = \mathbf{x}^{(t)}, \mathbf{A} = \mathbf{a})$, $t = L_3, ..., L_2$ can be used to measure precision and to construct confidence intervals for π_i, respectively. Obviously, instead of analyzing a single dependent Markov chain, it is also possible to generate $L_2 - L_3 + 1$ independent Markov chains of length L_3, where only the last iteration of each chain is saved for the estimation.

2.7 Simulation study

The simulations in this section are conducted to assess the benefit and the quality of the estimation procedures given in Sections 2.4-2.6. We run all simulations with MATLAB. We choose the true

parameter $\pi = (0.3, 0.4, 0.3)$, which may represent the proportions of persons in certain income classes, and $(\mathbb{P}(W = 1), ..., \mathbb{P}(W = 3)) = (2/3, 1/6, 1/6)$, where W represents a nonsensitive auxiliary characteristic. Groenitz (2012) presents ways to construct a W for a given distribution and shows that the above distribution of W provides a medium degree of privacy protection. The design matrix is then given by

$$C = \begin{pmatrix} c_1 & c_2 & c_3 \\ c_2 & c_3 & c_1 \\ c_3 & c_1 & c_2 \end{pmatrix} = \begin{pmatrix} 2/3 & 1/6 & 1/6 \\ 1/6 & 1/6 & 2/3 \\ 1/6 & 2/3 & 1/6 \end{pmatrix}.$$

We consider sample sizes $n \in \{100, 300\}$, the confidence level $1 - \alpha = 0.95$, and three Dirichlet(δ) prior distributions whose scatter plots appear in Figure 1. In particular, we study $\delta^{(1)} = (1, 1, 1)$, $\delta^{(2)} = (2.8, 4.3, 2.9)$, and $\delta^{(3)} = (7, 10.75, 7.25)$. The first is the noninformative prior, the second and third are informative priors. Both informative priors correspond to an investigator's guess $(\hat{\pi}_1^{(p)}, \hat{\pi}_2^{(p)}, \hat{\pi}_3^{(p)}) = (0.28, 0.43, 0.29)$ with $d^{(2)} = 10$ and $d^{(3)} = 25$, i.e, prior three indicates a larger certainty about the guess than prior two. In other words, prior three is more informative than prior two.

The simulation procedure is as follows. We draw 1000 samples of size n. In each sample, we calculate the posterior mode and apply parameter simulation (PS), multiple imputation (MI), and Rao-Blackwellization (RB) according to Sections 2.4-2.6 to calculate estimates and confidence intervals for the true π_i. The estimation quality is evaluated by the average estimate for π_i, the empirical MSE of the estimates for π_i, the empirical width, and the empirical coverage probability (CP) of the confidence intervals for π_i. The simulation results for PS, MI, and RB based on a single dependent Markov chain of length 1000 with burn-in period $t = 1, ..., 500$ are reported in Table 1 in the appendix.
For each of the methods PS, MI, and RB and for both considered sample sizes, we recognize that the average estimates are always close to the true proportions. The simulated MSEs and the widths of the CIs decrease as the prior becomes more informative. Additionally, we observe the tendency that the more informative the prior, the higher the coverage probabilities.

Reduced MSEs and shorter CIs are the effects caused by increasing the sample size.

Comparing the MSEs of the estimates for π_i, we find that RB and PS have nearly identical values, whereas MI shows the largest MSEs. The confidence widths of RB are smaller than the widths of MI, and PS delivers the widest CIs. However, RB has the lowest and PS has clearly the highest CPs. Due to the MSE results and the highest CPs, we evaluate PS to be the best method.

For comparison, we calculate the maximum likelihood estimates (MLEs) for each 1000 samples of size $n = 300$ and $n = 100$ and compute Bootstrap CIs (without normality assumption) for the π_i for each sample from $B = 2000$ Bootstrap replications, see Groenitz (2012), Section 3.2 and 3.3. The average ML estimates (see Table 3 in the appendix) are close to the true proportions. Consider $n = 300$ first. For the uniform prior ($\delta^{(1)}$), the CI widths and CPs for PS are slightly smaller than for ML. The MSEs of PS and ML are close to each other. The reason is that the posterior variance is a consistent estimate for the large sample variance of the ML estimator (see e.g., Little and Rubin (2002), Section 9.2.4). Parameter simulation with the informative prior with $\delta^{(2)}$ reduces the MSEs provided by ML by up to approximately 20%, and the more informative prior with $\delta^{(3)}$ leads to a reduction by approximately 40%.
We next examine $n = 100$. We notice that PS with the noninformative prior has smaller MSEs than ML. Moreover, we point out that PS with $\delta^{(2)}$ and $\delta^{(3)}$ decreases the MSEs of ML by approximately 40% and 75%, respectively. The widths of the CIs for π_i decrease by approximately 15% for $\delta^{(2)}$ and 30% for $\delta^{(3)}$ by using PS instead of ML.
For both informative priors and both sample sizes, there is a tendency that the CPs of PS are larger than the CPs of ML and overachieve the 95% level.
The estimates generated by PS are posterior means. On average, these posterior means are close to

the posterior modes (see appendix, Table 4). The MSEs of the posterior means and modes are quite similar for $n = 300$. In the case $n = 100$, the posterior modes provide a bit higher MSEs. We remark that the posterior mode for the uniform prior equals the MLE, if both are calculated from the same sample. This explains that the average MLEs and posterior means as well as the corresponding MSEs in Tables 3 and 4 are close to each other.

We also have conducted simulations in which the Bayes estimates were computed with the help of independent Markov chains. In particular, for each of 1000 simulated samples, we have calculated the PS, MI, and RB estimates from 500 independent chains of length 501, where only the last iteration of each chain is saved for the estimation. The simulation results are provided in Table 2. We discover that the above statements regarding estimates based on a single MC remain valid for the estimation with independent chains.

In sum, we emphasize that the estimation accuracy can be significantly improved by using Bayesian methods when prior information is available.

3 Common approach for Bayes estimation in privacy-protecting survey designs

Studying the calculations to obtain posterior modes and estimates based on parameter simulation, multiple imputation, and Rao-Blackwellization in Section 2, we observe that the design matrix C is the only component of the diagonal model that influences these calculations. Let us now consider an arbitrary PP design for $X \in \{1, ..., k\}$ with k_A possible scrambled answers and S required samples (in the DM, k_A equals k and $S = 1$). For each sample, we then have one design matrix. In the sequel, we restrict to PP designs whose design matrices do not contain nuisance terms, i.e., unknown parameters. For such a design, the only model component that is needed to compute Bayes estimates is the set of design matrices. That is, all relevant model information is stored in the design matrices - it does not matter whether we consider a RR or NRR method, moreover, the concrete answer scheme is irrelevant. Hence, most PP models for categorical X can be handled by a common approach. This fact has not been addressed in existing papers about Bayesian inference in PP models.
In Subsection 3.1, we give the design matrices for some PP models. Subsequently, in Subsection 3.2, we develop a general framework for Bayes estimation in PP designs for categorical X. Here, we generalize the calculations from Section 2 in order to cover many PP designs including certain multi-stage and multi-sample techniques.

3.1 Other privacy-protecting designs for categorical sensitive variables

We consider PP designs (i.e., RR or NRR models) for categorical sensitive variables $X \in \{1, ..., k\}$ with k_A possible answers (coded with $1, ..., k_A$) and S required samples. The complete data, i.e., the union of missing and observed data, are given by the vectors $(X_{sj}, A_{sj})_{sj}$ where X_{sj} and A_{sj} denote the X-value and the scrambled answer of respondent j in sample s, respectively ($s = 1, ..., S; j = 1, ..., n_s$). We demand the following conditions:

(M1) The $n = n_1 + ... + n_S$ vectors (X_{sj}, A_{sj}) are independent. Further, for $s = 1, ..., S$, the n_s vectors $(X_{s1}, A_{s1}), ..., (X_{s,n_s}, A_{s,n_s})$ are identically distributed, and $X_{sj} \sim X$ for all indices s, j.

(M2) The $k_A \times k$ matrices of conditional probabilities $C_s = [C_s(i, j)]_{ij} = [\mathbb{P}(A_{s1} = i \,|\, X_{s1} = j)]_{ij}$ have known entries ($s = 1, ..., S$).

Assumption (M1) means that the design needs S independent simple random samples with replacement (SRSWR) where the distribution of the scrambled answer is allowed to alter in different samples. We call the matrices C_s "design matrices". We next provide some examples of PP survey techniques, for which (M1)-(M2) are satisfied. All PP designs considered in the sequel are assumed to be applied to a SRSWR (for $S = 1$) respectively $S \geq 1$ independent SRSWR.

The RR model by Warner (1965) considers $X \in \{1, 2\}$ and needs one SRSWR. Each respondent draws and answers one of the questions "Do you have $X = 1$?" and "Do you have $X = 2$?". The first question is drawn with known probability c. The possible answers are "yes" and "no" (coded with 1 and 2). Then, the rows of $C = C_1$ are known and given by $(c, 1 - c)$ and $(1 - c, c)$.

The RR design by Abul-Ela, Greenberg, Horvitz (1967) is applicable to $X \in \{1, ..., k\}$, $k \geq 2$, and needs $S = k - 1$ independent samples (each sample is a SRSWR). The interviewees select and answer one of the k questions "Do you have $X = j$?" $(j = 1, ..., k)$. The probability c_{sj} $(s = 1, ..., k - 1; j = 1, ..., k)$ that question j is selected in sample s is determined by the RD and is known. Coding "yes" and "no" by 1 and 2 results in the $2 \times k$ matrices C_s having the j-th column equal to $(c_{sj},\ 1-c_{sj})^T$ $(s = 1, ..., k-1)$.

The unrelated question model (UQM) - see Horvitz et al. (1967) and Greenberg et al. (1969) - is constructed for a sensitive $X \in \{1, 2\}$. According to the result of a random experiment, each interviewee answers either "Do you have $X = 1$?" or "Do you have $Y = 1$?" where $Y \in \{1, 2\}$ is an unrelated nonsensitive variable. Let c be the known probability that the first question is selected, and assume $\phi = \mathbb{P}(Y = 1)$ to be known. Then, the UQM requires a single SRSWR, and we have $C = C_1$ with rows $(c + (1 - c)\phi,\ (1 - c)\phi)$ and $((1 - c)(1 - \phi),\ (1 - c)(1 - \phi) + c)$. If the distribution of Y is unknown, the UQM needs two independent SRSWR. In this case, we can define the new variable

$$\tilde{X} \in \{1, ..., 4\} \tag{8}$$

that attains the values 1, 2, 3, 4 if (X, Y) attains $(1, 1)$, $(1, 2)$, $(2, 1)$, $(2, 2)$, respectively. This $\tilde{X}$ plays the role of X from (M1) and (M2). Let c_{s1} be the known probability that question 1 is selected in sample s. It follows that C_s has the rows $(1,\ c_{s1},\ 1 - c_{s1},\ 0)$ and $(0,\ 1 - c_{s1},\ c_{s1},\ 1)$.

Omitting details, we also can fulfill (M1)-(M2) for the RR methods for $X \in \{1, ..., k\}$ $(k \geq 2)$ suggested by Eriksson (1973), and Liu et al. (1975).
The two-stage RR design by Mangat and Singh (1990) considers $X \in \{1, 2\}$. In the first stage, each respondent conducts a random experiment that decides whether the question "Do you have $X = 1$?" must be answered or whether the respondent has to go to stage two. In stage two, another random experiment must be accomplished by the interviewee. According to its outcome, either the question "Do you have $X = 1$?" or "Do you have $X = 2$?" must be answered. This model needs one SRSWR, and $C = C_1$ has the known rows $(T + (1 - T)c, (1 - T)(1 - c))$ and $((1 - T)(1 - c), T + (1 - T)c)$, where T is the probability that the experiment in stage one decides that the question must be answered and c is the probability of drawing the first question in stage two.
Omitting certain details again, for the RR model by Mangat (1994), (M1)-(M2) are fulfilled, where $k_A = 2$, $S = 1$, and $C = C_1$ with rows $(1, 1 - c)$ and $(0, c)$ for a $c \in (0, 1)$.

Quatember (2009) presents a standardized RR model for $X \in \{1, 2\}$ and explains that 16 survey designs are special cases of his model. In this standardized design, each interviewee draws randomly one of the five instructions:

1:	Answer "Do you have $X = 1$?"	2:	Answer "Do you have $X = 2$?"		
3:	Answer "Do you have $Y = 1$?"	4:	Say "yes"	5:	Say "no"

Here, $Y \in \{1, 2\}$ is a nonsensitive characteristic. Let us consider a single SRSWR, set $\phi = \mathbb{P}(Y = 1)$, and define c_i to be the probability that instruction i is drawn. Coding answers "yes" and

"no" with 1 and 2 yields the 2×2 design matrix with rows $(c_1 + c_3\phi + c_4,\ c_2 + c_3\phi + c_4)$ and $(c_2 + c_3(1-\phi) + c_5,\ c_1 + c_3(1-\phi) + c_5)$ and (M1)-(M2) are fulfilled.

The properties (M1)-(M2) are also satisfied for the following NRR models: the hidden sensitivity model by Tian et al. (2007), the crosswise and triangular model by Yu et al. (2008), and the multicategory model by Tang et al. (2009). For instance, Tang et al. (2009) consider $X \in \{1, ..., k\}$, $k \geq 2$. The respondent's answer depends on the value of X and on the value of a nonsensitive auxiliary variable $W \in \{1, ..., k\}$, which is independent of X and possesses a known distribution (e.g., W may describe the period of the birthday). If $X = 1$, an answer equal to the value of W is required. For $X = i$, the response i ($i = 2, ..., k$) must be given. The design needs a single SRSWR. The first column of the $k \times k$ matrix $C = C_1$ equals $(P(W = 1), ..., \mathbb{P}(W = k))^T$, and column i ($i = 2, ..., k$) is a vector having entry i equal to 1 and all other entries equal to 0.

We finish this section with a model that violates (M2): the two-trial UQM by Horvitz et al. (1967) is for $X \in \{1, 2\}$ and needs $S = 2$ independent SRSWR. Each respondent selects one of the questions "Do you have $X = 1$?" or "Do you have $Y = 1$?" with the help of a random experiment (Y is again an unrelated variable). Subsequently, the selection is repeated. The possible answers are 1=("yes", "yes"), 2=("yes", "no"), 3=("no", "yes"), 4=("no", "no"). The distribution of Y is unknown, and independence between X and Y is assumed. Then, we have

$$C_s = \begin{pmatrix} c_{s1}^2 + 2c_{s1}c_{s2}\phi + c_{s2}^2\phi & c_{s2}^2\phi \\ c_{s1}c_{s2}(1-\phi) & c_{s1}c_{s2}\phi \\ c_{s1}c_{s2}(1-\phi) & c_{s1}c_{s2}\phi \\ c_{s2}^2(1-\phi) & c_{s1}^2 + 2c_{s1}c_{s2}(1-\phi) + c_{s2}^2(1-\phi) \end{pmatrix}$$

with $s \in \{1, 2\}$, where $\phi = \mathbb{P}(Y = 1)$, c_{s1} is the known probability that question 1 is selected in sample s, and $c_{s2} = 1 - c_{s1}$. Since ϕ is unknown, (M2) does not hold. A possible remedy is to abandon the independence assumption for X and Y and to consider $\tilde{X}$ from (8) again. $\tilde{X}$ plays the role of X in (M1)-(M2) with

$$C_s = \begin{pmatrix} 1 & c_{s1}^2 & c_{s2}^2 & 0 \\ 0 & c_{s1}c_{s2} & c_{s1}c_{s2} & 0 \\ 0 & c_{s1}c_{s2} & c_{s1}c_{s2} & 0 \\ 0 & c_{s2}^2 & c_{s1}^2 & 1 \end{pmatrix},$$

where $s \in \{1, 2\}$. This version of the two-trial UQM, which can be found in Bourke and Moran (1988), Section 2, satisfies (M1)-(M2).

3.2 Bayes estimation in PP models

The calculations from Section 2 can be generalized to arbitrary randomized response and nonrandomized response survey techniques with (M1)-(M2). For such a model, the missing data $\mathbf{X}$ and observed data $\mathbf{A}$ are given by $(X_{sj})_{sj}$ and $(A_{sj})_{sj}$, respectively ($s = 1, ..., S$; $j = 1, ..., n_s$). Set for $x_{sj} \in \{1, ..., k\}$ and $a_{sj} = \{1, ..., k_A\}$

$$f_{\mathbf{X},\mathbf{A}\,|\,\Pi}(\mathbf{x}, \mathbf{a}\,|\,\pi) = \prod_{s=1}^{S} \prod_{j=1}^{n_s} C_s(a_{sj}, x_{sj}) \cdot \pi_{x_{sj}},$$

where the C_s are the design matrices of the PP model and $\mathbf{x} = (x_{sj})_{sj}$, $\mathbf{a} = (a_{sj})_{sj}$. Accordingly, we have

$$f_{\mathbf{X}\,|\,\mathbf{A},\Pi}(\mathbf{x}\,|\,\mathbf{a}, \pi) = \prod_{s=1}^{S} \prod_{j=1}^{n_s} \frac{C_s(a_{sj}, x_{sj}) \cdot \pi_{x_{sj}}}{f_{A_{sj}\,|\,\Pi}(a_{sj}\,|\,\pi)},$$

where $f_{A_{sj}\,|\,\Pi}(\alpha\,|\,\pi)$ is the entry number $\alpha \in \{1, ..., k_A\}$ of vector $C_s \cdot (\pi_1, ..., \pi_k)^T$. As in Section 2, we focus on Dirichlet prior distributions.

To calculate the posterior mode in a PP design with (M1)-(M2), (6) becomes

$$Q(\pi \,|\, \pi^{(t)}) = \sum_{s=1}^{S} \sum_{i=1}^{k} \hat{m}_{si}^{(t)} \cdot \log \pi_i \text{ and } Q^*(\pi \,|\, \pi^{(t)}) = \sum_{i=1}^{k} \left(\delta_i - 1 + \sum_{s=1}^{S} \hat{m}_{si}^{(t)} \right) \cdot \log \pi_i$$

with $\hat{m}_{si}^{(t)} = \sum_{j=1}^{k_A} n_{sj} \cdot \pi_i^{(t)} \cdot C_s(j,i) / f_{A_{s1} \,|\, \Pi}(j \,|\, \pi^{(t)})$, where n_{sj} is the number of respondents in sample s giving answer j. The term $\hat{m}_{si}^{(t)}$ is equal to the sum of the i-th column of the $k_A \times k$ matrix

$$C_s \,.^* \left[\left[\tilde{n}_s^T ./ \, \lambda_s(\pi^{(t)}) \right] \cdot (\pi_1^{(t)}, ..., \pi_k^{(t)}) \right] \text{ with}$$

$$\tilde{n}_s = (n_{s1}, ..., n_{sk_A}) \text{ and } \lambda_s(\pi^{(t)}) = (f_{A_{s1} \,|\, \Pi}(1 \,|\, \pi^{(t)}), ..., f_{A_{s1} \,|\, \Pi}(k_A \,|\, \pi^{(t)}))^T.$$

Maximization of $Q^*(\cdot \,|\, \pi^{(t)})$ results in $\pi_i^{(t+1)} = (\delta_i - 1 + \sum_{s=1}^{S} \hat{m}_{si}^{(t)}) / (n - k + \delta_1 + ... + \delta_k)$.

To conduct parameter simulation and to obtain multiple imputations, data augmentation for a general privacy-protecting survey design proceeds as follows:
I step: It suffices to simulate the number of sample units with $X = j$. Let $m_s^{(t)}(i,j)$ be the in iteration t simulated number of persons who have X-value j among the persons who give answer i in sample s. Draw

$$(m_s^{(t)}(i,1), ..., m_s^{(t)}(i,k)) \sim Multinomial(n_{si}, \gamma_{s,i}^{(t)}).$$

The vector $\gamma_{s,i}^{(t)}$ contains the cell probabilities and is defined to be the i-th row of the $k_A \times k$ matrix

$$C_s \,.^* \left[\left[(1, \cdots, 1)^T ./ \, \lambda_s(\pi^{(t-1)}) \right] \cdot \left(\pi_1^{(t-1)}, ..., \pi_k^{(t-1)} \right) \right],$$

where

$$\lambda_s(\pi^{(t-1)}) = (f_{A_{s1} \,|\, \Pi}(1 \,|\, \pi^{(t-1)}), ..., f_{A_{s1} \,|\, \Pi}(k_A \,|\, \pi^{(t-1)}))^T.$$

Obviously, the cell probabilities depend (apart from the parameters of the preceding iteration) only on the design matrices. The desired number of persons having $X = j$ in iteration t is then $m_j^{(t)} = \sum_{s=1}^{S} \sum_{i=1}^{k_A} m_s^{(t)}(i,j)$.

P step: Draw a new parameter $(\pi_1^{(t)}, ..., \pi_{k-1}^{(t)})^T$ from $f_{\Pi \,|\, \mathbf{X}, \mathbf{A}}(\cdot \,|\, \mathbf{x}^{(t)}, \mathbf{a})$, a density corresponding to the $Di(m_1^{(t)} + \delta_1, ..., m_k^{(t)} + \delta_k)$ distribution.

Rao-Blackwellized estimates for a general PP design can be obtained analogously to Subsection 2.6 by averaging conditional expectations. In particular, the estimate is given by

$$\tilde{s} = (L_2 - L_3 + 1)^{-1} \sum_{t=L_3}^{L_2} \mathbb{E}(\Pi \,|\, \mathbf{X} = \mathbf{x}^{(t)}, \mathbf{A} = \mathbf{a}).$$

with (compare P step of data augmentation above)

$$\mathbb{E}(\Pi \,|\, \mathbf{X} = \mathbf{x}^{(t)}, \mathbf{A} = \mathbf{a}) = \frac{(m_1^{(t)} + \delta_1, ..., m_{k-1}^{(t)} + \delta_{k-1})^T}{(n + \delta_1 + ... + \delta_k)},$$

where $m_j^{(t)}$ is again the simulated count of persons having $X = j$ in iteration t.

4 Summary

Survey concepts that protect the respondents' privacy are important to obtain reliable data on sensitive characteristics. To exploit prior information on the distribution of the sensitive variable, the application of Bayesian methods is appealing. In this paper, we have developed a Bayesian extension of the privacy-protecting, nonrandomized diagonal model survey technique by Groenitz (2012). We illustrated in simulations that precision can be significantly improved by incorporating available prior information into the estimation. In the second part of this paper, we found that for any privacy-protecting survey design dealing with categorical sensitive characteristics, all relevant model information is stored in the design matrices. For this reason, we were able to present the Bayes inference for privacy-protecting models in a general framework that covers a lot of randomized and nonrandomized response methods.

References

[1] Abul-Ela, A.A., Greenberg, B.G., Horvitz, D.G.: A Multi-Proportions Randomized Response Model. Journal of the American Statistical Association 62, 990-1008 (1967)

[2] Barabesi, L., Marcheselli, M.: Bayesian estimation of proportion and sensitivity level in randomized response procedures. Metrika 72, 75-88 (2010)

[3] Bourke, P.D., Moran, M.A.: Estimating Proportions From Randomized Response Data Using the EM Algorithm. Journal of the American Statistical Association 83, 964-968 (1988)

[4] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B 39, 1-38 (1977)

[5] Eriksson, S.A.: A New Model for Randomized Response. International Statistical Review 41, 101-113 (1973)

[6] Gentle, J.E.: Random Number Generation and Monte Carlo Methods. Springer (1998)

[7] Greenberg, B.G., Abul-Ela, A.A., Simmons, W.R., Horvitz, D.G.: The Unrelated Question Randomized Response Model: Theoretical Framework. Journal of the American Statistical Association 64, 520-539 (1969)

[8] Groenitz, H.: A New Privacy-Protecting Survey Design for Multichotomous Sensitive Variables. Metrika, DOI: 10.1007/s00184-012-0406-8 (2012).

[9] Horvitz, D.G., Shah, B.V., Simmons, W.R.: The Unrelated Question Randomized Response Model. Proceedings of the Social Statistics Section, American Statistical Association, 65-72 (1967)

[10] Hussain, Z., Cheema, S.A., Zafar, S.: Extension of Mangat Randomized Response Model. International Journal of Business and Social Science 2, 261-266 (2011)

[11] Kim, J.M., Tebbs, J.M., An, S.W.: Extensions of Mangat's randomized-response model. Journal of Statistical Planning and Inference 136, 1554-1567 (2006)

[12] Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley (2002)

[13] Liu, P.T., Chow, L.P., Mosley, W.H.: Use of the Randomized Response Technique With a New Randomizing Device. Journal of the American Statistical Association 70, 329-332 (1975)

[14] Mangat, N.S.: An Improved Randomized Response Strategy. Journal of the Royal Statistical Society B 56, 93-95 (1994)

[15] Mangat, N.S., Singh, R.: An Alternative Randomized Response Procedure. Biometrika 77, 439-442 (1990)

[16] Migon, H.S., Tachibana, V.M.: Bayesian approximations in randomized response model. Computational Statistics & Data Analysis 24, 401-409 (1997)

[17] O'Hagan, A.: Bayes Linear Estimators for Randomized Response Models. Journal of the American Statistical Association 82, 207-214 (1987)

[18] Quatember, A.: A standardization of randomized response strategies. Statistics Canada, Survey Methodology 35, 143-152 (2009)

[19] Schafer, J.L.: Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC (2000)

[20] Tan, M.T., Tian, G.L., Tang, M.L.: Sample Surveys with Sensitive Questions: A Nonrandomized Response Approach. The American Statistician 63, 9-16 (2009)

[21] Tanner, M.A., Wong, W.H.: The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association 82, 528-540 (1987)

[22] Tang, M.L., Tian G.L., Tang, N.S., Liu, Z.: A new non-randomized multi-category response model for surveys with a single sensitive question: Design and analysis. Journal of the Korean Statistical Society 38, 339-349 (2009)

[23] Tian, G.L., Yu, J.W., Tang, M.L., Geng, Z.: A new non-randomized model for analysing sensitive questions with binary outcomes. Statistics in Medicine 26, 4238-4252 (2007)

[24] Tian, G.L., Yuen, K.C., Tang, M.L., Tan, M.T.: Bayesian non-randomized response models for surveys with sensitive questions. Statistics and its interface 2, 13-25 (2009)

[25] Unnikrishnan, N.K., Kunte, S.: Bayesian analysis for randomized response models. The Indian Journal of Statistics 61, Series B, 422-432 (1999)

[26] Warner, S.L.: Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association 60, 63-69 (1965)

[27] Winkler, R.L., Franklin, L.A.: Warner's Randomized Response Model: A Bayesian Approach. Journal of the American Statistical Association 74, 207-214 (1979)

[28] Yu, J.W., Tian, G.L., Tang, M.L.: Two new models for survey sampling with sensitive characteristic: design and analysis. Metrika 67, 251-263 (2008)

A Appendix: Simulation Outputs

This appendix contains the simulation results described in Section 2.7.

		$n = 300$ - estimation based on a single Markov chain											
		Parameter simulation				Multiple imputation				Rao-Blackwellization			
		av.est.	MSE	width	CP	av.est.	MSE	width	CP	av.est.	MSE	width	CP
$\delta^{(1)}$	π_1	0.2986	0.0027	0.2071	0.9540	0.2982	0.0028	0.1827	0.9300	0.2986	0.0027	0.1809	0.9260
	π_2	0.3972	0.0029	0.2140	0.9410	0.3979	0.0030	0.1873	0.9140	0.3972	0.0029	0.1854	0.9070
	π_3	0.3043	0.0028	0.2075	0.9470	0.3039	0.0028	0.1830	0.9180	0.3042	0.0028	0.1812	0.9140
$\delta^{(2)}$	π_1	0.2969	0.0022	0.1970	0.9610	0.2974	0.0023	0.1760	0.9250	0.2969	0.0022	0.1704	0.9190
	π_2	0.4070	0.0025	0.2047	0.9610	0.4063	0.0027	0.1812	0.9240	0.4070	0.0025	0.1753	0.9180
	π_3	0.2961	0.0027	0.1971	0.9330	0.2963	0.0028	0.1758	0.9130	0.2961	0.0026	0.1701	0.9030
$\delta^{(3)}$	π_1	0.2942	0.0017	0.1799	0.9720	0.2954	0.0019	0.1645	0.9470	0.2942	0.0016	0.1518	0.9380
	π_2	0.4077	0.0018	0.1886	0.9740	0.4058	0.0021	0.1700	0.9450	0.4076	0.0018	0.1569	0.9420
	π_3	0.2981	0.0015	0.1803	0.9740	0.2988	0.0018	0.1644	0.9490	0.2981	0.0015	0.1518	0.9450
		$n = 100$ - estimation based on a single Markov chain											
		Parameter simulation				Multiple imputation				Rao-Blackwellization			
		av.est.	MSE	width	CP	av.est.	MSE	width	CP	av.est.	MSE	width	CP
$\delta^{(1)}$	π_1	0.2956	0.0078	0.3460	0.9470	0.2945	0.0083	0.3142	0.9140	0.2957	0.0078	0.3050	0.9030
	π_2	0.3985	0.0082	0.3625	0.9450	0.4004	0.0087	0.3249	0.9170	0.3985	0.0082	0.3154	0.9060
	π_3	0.3059	0.0078	0.3477	0.9480	0.3050	0.0082	0.3154	0.9220	0.3058	0.0077	0.3063	0.9100
$\delta^{(2)}$	π_1	0.2974	0.0046	0.3047	0.9670	0.2991	0.0056	0.2836	0.9340	0.2974	0.0046	0.2578	0.9290
	π_2	0.4090	0.0053	0.3189	0.9720	0.4070	0.0064	0.2923	0.9400	0.4091	0.0053	0.2657	0.9300
	π_3	0.2936	0.0046	0.3027	0.9700	0.2939	0.0056	0.2815	0.9450	0.2936	0.0046	0.2559	0.9350
$\delta^{(3)}$	π_1	0.2898	0.0023	0.2514	0.9900	0.2922	0.0035	0.2476	0.9680	0.2897	0.0023	0.1981	0.9570
	π_2	0.4151	0.0026	0.2673	0.9880	0.4115	0.0039	0.2595	0.9660	0.4152	0.0026	0.2076	0.9510
	π_3	0.2951	0.0021	0.2514	0.9960	0.2963	0.0033	0.2470	0.9740	0.2950	0.0021	0.1976	0.9580

Table 1: *Simulation results for PS, MI, RB based on a single Markov chain. The performance of the estimation strategies is assessed in terms of the average estimate for π_i, the simulated MSE of the estimates for π_i, the empirical width and coverage probability of the confidence intervals for π_i ($\alpha = 5\%$). The true proportions are given by* (0.3, 0.4, 0.3).

$n = 300$ - estimation based on independent Markov chains												
		Parameter simulation				Multiple imputation				Rao-Blackwellization			
		av.est.	MSE	width	CP	av.est.	MSE	width	CP	av.est.	MSE	width	CP
$\delta^{(1)}$	π_1	0.2971	0.0027	0.2080	0.9550	0.2968	0.0028	0.1837	0.9200	0.2971	0.0027	0.1819	0.9110
	π_2	0.4004	0.0032	0.2155	0.9490	0.4010	0.0032	0.1883	0.9140	0.4004	0.0032	0.1864	0.9110
	π_3	0.3024	0.0029	0.2083	0.9440	0.3022	0.0030	0.1838	0.9080	0.3025	0.0029	0.1819	0.9030
$\delta^{(2)}$	π_1	0.2963	0.0024	0.1983	0.9490	0.2969	0.0025	0.1767	0.9180	0.2963	0.0024	0.1710	0.9120
	π_2	0.4074	0.0026	0.2058	0.9510	0.4066	0.0028	0.1818	0.9140	0.4074	0.0026	0.1760	0.9090
	π_3	0.2963	0.0022	0.1982	0.9570	0.2965	0.0024	0.1770	0.9210	0.2963	0.0022	0.1713	0.9150
$\delta^{(3)}$	π_1	0.2944	0.0017	0.1814	0.9690	0.2955	0.0019	0.1653	0.9360	0.2943	0.0017	0.1526	0.9310
	π_2	0.4091	0.0018	0.1899	0.9740	0.4074	0.0021	0.1712	0.9370	0.4091	0.0018	0.1580	0.9280
	π_3	0.2965	0.0017	0.1811	0.9650	0.2971	0.0020	0.1653	0.9310	0.2965	0.0017	0.1526	0.9290
$n = 100$ - estimation based on independent Markov chains													
		Parameter simulation				Multiple imputation				Rao-Blackwellization			
		av.est.	MSE	width	CP	av.est.	MSE	width	CP	av.est.	MSE	width	CP
$\delta^{(1)}$	π_1	0.3000	0.0071	0.3504	0.9590	0.2991	0.0076	0.3186	0.9350	0.3001	0.0071	0.3094	0.9280
	π_2	0.3956	0.0082	0.3645	0.9520	0.3975	0.0087	0.3276	0.9300	0.3957	0.0083	0.3180	0.9140
	π_3	0.3043	0.0085	0.3499	0.9420	0.3034	0.0089	0.3171	0.9080	0.3043	0.0084	0.3078	0.8990
$\delta^{(2)}$	π_1	0.2911	0.0047	0.3040	0.9710	0.2921	0.0057	0.2823	0.9360	0.2910	0.0047	0.2566	0.9240
	π_2	0.4080	0.0049	0.3212	0.9780	0.4059	0.0059	0.2942	0.9520	0.4081	0.0049	0.2675	0.9430
	π_3	0.3009	0.0045	0.3058	0.9820	0.3021	0.0054	0.2841	0.9510	0.3010	0.0045	0.2583	0.9380
$\delta^{(3)}$	π_1	0.2880	0.0022	0.2513	0.9980	0.2900	0.0032	0.2478	0.9800	0.2880	0.0022	0.1982	0.9680
	π_2	0.4166	0.0028	0.2683	0.9910	0.4133	0.0041	0.2602	0.9700	0.4166	0.0028	0.2081	0.9600
	π_3	0.2954	0.0022	0.2528	0.9930	0.2968	0.0034	0.2486	0.9680	0.2954	0.0022	0.1988	0.9560

Table 2: *Simulation results for PS, MI, RB based on independent Markov chains. The performance of the estimation strategies is assessed in terms of the average estimate for π_i, the simulated MSE of the estimates for π_i, the empirical width and coverage probability of the confidence intervals for π_i ($\alpha = 5\%$). The true proportions are given by* (0.3, 0.4, 0.3).

ML estimation for $n = 300$				
	av.est.	MSE	width	coverage
π_1	0.2996	0.0028	0.2097	0.9580
π_2	0.4008	0.0030	0.2174	0.9510
π_3	0.2996	0.0028	0.2102	0.9470
ML estimation for $n = 100$				
π_1	0.3024	0.0084	0.3587	0.9580
π_2	0.4008	0.0094	0.3735	0.9510
π_3	0.2968	0.0083	0.3584	0.9500

Table 3: *This table contains the simulation results for the ML estimation based on 1000 samples. Average ML estimates for π_i, empirical MSEs for the ML estimates as well as empirical widths and coverage probabilities for Bootstrap CIs ($\alpha = 5\%$) reported. The true proportions are given by* (0.3, 0.4, 0.3).

		Posterior modes			
		$n = 300$		$n = 100$	
		av. est.	MSE	av. est.	MSE
$\delta^{(1)}$	π_1	0.2979	0.0027	0.2942	0.0086
	π_2	0.3982	0.0030	0.4013	0.0089
	π_3	0.3040	0.0028	0.3045	0.0084
$\delta^{(2)}$	π_1	0.2964	0.0022	0.2960	0.0052
	π_2	0.4080	0.0026	0.4126	0.0060
	π_3	0.2956	0.0027	0.2914	0.0052
$\delta^{(3)}$	π_1	0.2940	0.0017	0.2880	0.0026
	π_2	0.4085	0.0019	0.4186	0.0030
	π_3	0.2976	0.0016	0.2934	0.0024

Table 4: *Simulation results for the observed data posterior mode. The table reports the average posterior mode and the corresponding empirical MSE. The true proportions are given by* (0.3, 0.4, 0.3).

```
function [PS_stats, MI_stats, RB_stats, post_mode, Iter]=...
    Bayes_est(nn,C,L,de,al)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%↙
%%%%%%%%%%%

% Supplemental material for the manuscript
% Groenitz, H.: Using Prior Information in Privacy-Protecting
% Survey Designs for Categorical Sensitive Variables.

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%↙
%%%%%%%%%%%



% This function enables Bayesian estimation in randomized and
% nonrandomized response models for categorical sensitive variables.
% The number of required samples in the model is denoted with S, the
% sensitive variable has k categories, k_A different answers are possible.

% I  N  P  U  T:
% nn: S x k_A matrix; entry (s,j) is the number of respondents in sample s
% giving answer j

% C: S*k_A x k matrix; collects design matrices for the S samples one below
% the other. The matrix C must not contain unknown parameters.

% L is a vector [L(1) L(2) L(3)] with L(1): number of independent Markov chains
% generated by data augmentation, L(2): length of each Markov chain, the
% realizations from iteration L(3), L(3)+1,...,L(2) of  each chain are used for
% the estimation

% de: 1 x k parameter vector of the Dirichlet prior distribution

% al: 1-al is the required level of the Bayes confidence intervals

% O  U  T  P  U  T:
% The structure array PS_stats contains quantities that are calculated by
% parameter simulation (PS) and has the fields
% B_mean_PS, B_std_PS, B_CI_PS. Here, the k x 1 vectors B_mean_PS and
% B_std_PS contain the componentwise mean and standard deviation of the
% draws from the observed data posterior, respectively.
% B_CI_PS is a k x 2 matrix containing Bayes 1-al confidence intervals
% for the k unknown proportions

% Analogously, the structure array MI_stats possesses the fields B_mean_MI,
% B_std_MI, B_CI_MI, which are quantities calculated from multiple
% imputations. The structure array RB_stats has the fields  B_mean_RB,
% B_std_RB, B_CI_RB, which represent quantities derived by Rao-Blackwellization.

% Post_mode: Observed data posterior mode computed with EM algorithm
% Iter: Number of iterations of EM algorithm to calculate the posterior mode

%----------------------------------------------------------------------
% A more detailed description of this program including examples for its
% application is attached in the form of a pdf-file.
%↙
=======================================================================================↙
=
```

```
k=length(C(1,:)); S=length(nn(:,1)); k_A=length(nn(1,:));n=sum(sum(nn));

%Posterior mode via EM algorithm

pi1= ones(k,1)/k; % starting value
%E step: Calculate Q*(pi|pi^t)=Q(pi|pi^t)+log f(pi)
la=C*pi1;
M=sum( C.* ((reshape(nn',S*k_A,1)./ la) * pi1'),1) + de -1;
%Q*(pi|pi^t)= M * (log pi_1,....,log pi_k)'

%M step
pi2= M'/sum(M);
Iter=1;
while max(abs(pi2-pi1)) > 10^-8
Iter=Iter+1;
pi1=pi2;
%E step
la=C*pi1;
M=sum( C.* ((reshape(nn',S*k_A,1)./ la) * pi1'),1) + de -1;
% M step
pi2= M'/sum(M);

end
post_mode=pi2;


% Generate  Markov chains with the help of the data augmentation algorithm
q=L(2)-L(3)+1; PI=zeros(L(1)*q,k); IMP=PI; RB=PI;

for i=1:L(1) %i-th Markov chain
   pi=ones(k,1)/k; % starting value
   E_ps=zeros(L(2), k); E_m=E_ps; E_rb=E_ps;
   for j=1:L(2)
      %I step:
      la=C*pi;
      cp=C .* ( (1./la) * pi');
      cp=cp./   repmat(sum(cp,2),1,k);
      M=sum(mnrnd(reshape(nn',S*k_A,1) ,cp),1); % M is a row vector;
      E_m(j,:) =M;
      E_rb(j,:)=(M+de)/(n+sum(de));

      %P step: Draw from the Dirichlet distribution with param. (M+de)'
      Y=gamrnd((M+de)',ones(k,1));
      pi=Y/sum(Y); %k x 1 vector
      E_ps(j,:)= pi';
   end
   PI ( (i-1)*q + 1 : i*q , 1:k)= E_ps(L(3):L(2),:);
   IMP( (i-1)*q + 1 : i*q , 1:k)=  E_m(L(3):L(2),:);
   RB ( (i-1)*q + 1 : i*q , 1:k)= E_rb(L(3):L(2),:);

end

% PI contains draws from the observed data posterior distribution
% Begin evaluation of the matrix PI

B_mean_PS = mean(PI,1)'; %columnwise mean
B_std_PS  = std(PI,0,1)'; %"0": division by (sample size - 1); "1": columnwise std
B_CI_PS   =[quantile(PI,al/2); quantile(PI,1-al/2)]';
```

```
PS_stats=struct('B_mean_PS',B_mean_PS,'B_std_PS',B_std_PS,'B_CI_PS',B_CI_PS);
%quantile: columnwise empirical quantiles, returns a row vector

% IMP contains multiple imputations
PI_MI=IMP/n; % PI_MI contains estimates for the true proportions computed from IMP
B_mean_MI = mean(PI_MI,1)'; %columnwise mean
B_std_MI  = std(PI_MI,0,1)'; %"0": division by (sample size - 1); "1": columnwise std
B_CI_MI=[quantile(PI_MI,al/2); quantile(PI_MI,1-al/2)]';
MI_stats=struct('B_mean_MI',B_mean_MI,'B_std_MI',B_std_MI,'B_CI_MI',B_CI_MI);

% Estimates motivated by Rao-Blackwell Theorem
B_mean_RB = mean(RB,1)'; %columnwise mean
B_std_RB  = std(RB,0,1)'; %"0": division by (sample size - 1); "1": columnwise std
B_CI_RB=[quantile(RB,al/2); quantile(RB,1-al/2)]';
RB_stats=struct('B_mean_RB',B_mean_RB,'B_std_RB',B_std_RB,'B_CI_RB',B_CI_RB);

end
```

Using Prior Information in Privacy-Protecting Survey Designs for Categorical Sensitive Variables

-

Description of the MATLAB program `Bayes_est.m`

Heiko Groenitz*

The MATLAB program `Bayes_est.m` computes Bayesian estimates in privacy protecting (PP) survey designs for categorical sensitive variables whose design matrices do not contain unknown parameters. The number of required samples in the model is denoted with S, the sensitive variable has k categories (coded with $1, ..., k$) and k_A different scrambled answers (coded with 1,...,k_A) are possible. The program has the following input variables:

- `nn` is a $S \times k_A$ matrix; entry (s, j) is the number of respondents in sample s giving answer j.
- `C` represents a $S \cdot k_A \times k$ matrix that collects the design matrices for the S samples one below the other.
- `L` is a vector `[L(1) L(2) L(3)]` with `L(1)`: number of independent Markov chains generated by data augmentation and `L(2)`: length of each Markov chain. The realizations from iteration `L(3), L(3)+1,...,L(2)` of each chain are used for the estimation, the realizations from iteration `1,..., L(3)-1` are rejected.
- `de` is a $1 \times k$ parameter vector of the Dirichlet prior distribution.
- `al` is a real number such that 1-`al` describes the required level of the Bayes confidence intervals.

The output of `Bayes_est.m` delivers estimates based on parameter simulation, multiple imputation and Rao-Blackwellization as well as the observed data posterior mode. In particular, we have:

- Parameter simulation means that we draw from the posterior distribution of the parameters given the observed data. The $k \times 1$ vectors `B_mean_PS` and `B_std_PS` contain the componentwise mean and standard deviation of these draws, respectively. `B_CI_PS` is a $k \times 2$ matrix containing Bayes 1-`al` confidence intervals (CIs) for the k unknown proportions. These CIs are based on simulated `al`/2 and 1-`al`/2 posterior quantiles. The fields `B_mean_PS`, `B_std_PS` and `B_CI_PS` are collected in the structure array `PS_stats`.
- The structure array `MI_stats` possesses the fields `B_mean_MI`, `B_std_MI` and `B_CI_MI`, which are quantities calculated from multiple imputations. Each imputation results in one estimate for the unknown proportions. `B_mean_MI` is the average estimate and `B_std_MI` provides the componentwise standard deviation of these estimates. The i-th row of the $k \times 2$ matrix `B_CI_MI` gives a 1-`al` Bayes confidence interval for the proportion of individuals who possess outcome i of the sensitive variable.
- The structure array `RB_stats` has the fields `B_mean_RB`, `B_std_RB` and `B_CI_RB`, which represent quantities derived by Rao-Blackwellization. The $k \times 1$ vectors `B_mean_RB` and `B_std_RB` provide the componentwise mean and standard deviation of the `L(1)·(L(2)-L(3)+1)` conditional expectations

  $$\mathbb{E}(\Pi \,|\, \mathbf{X} = \mathbf{x}^{(t)}, \mathbf{A} = \mathbf{a})$$

  that appear in the section about estimates motivated by the Rao-Blackwell theorem in the paper. The first (second) column of the $k \times 2$ matrix `B_CI_RB` contains the simulated `al`/2 (1-`al`/2) quantiles of the above mentioned `L(1)·(L(2)-L(3)+1)` conditional expectations (componentwise quantiles). That is, the i-th row of `B_CI_RB` provides a 1-`al` Bayes CI for the true proportion of units in the population having outcome i of the sensitive variable.
- `post_mode` is the observed data posterior mode computed with the EM algorithm.
- `Iter` is the number of iterations of the EM algorithm for the calculation of the posterior mode.

*Philipps-University Marburg, Department for Statistics (Faculty 02), Universitätsstraße 25, 35032 Marburg, Germany (e-mail: groenitz@staff.uni-marburg.de).

In the sequel, we consider concrete examples for the application of the program `Bayes_est.m`. Details of the considered PP designs can be found in the paper.

Example 1: Nonrandomized multi-category (MC) model by Tang et al. (2009)

Tang et al. (2009) present an illustrative example for their nonrandomized MC model. According to their data, we set

```
nn=[59 97 82 76 81];
c=[0.2 0.2 0.2 0.2 0.2]; k=length(c);
C=zeros(k,k); C(:,1)=c; C(2:k,2:k)=eye(k-1);
de=[1 1 1 1 1]; al=0.05; L=[1 40000 20001];
[PS_stats, MI_stats, RB_stats, post_mode, Iter] = Bayes_est(nn,C,L,de,al)
```

That is, the uniform prior is considered and data augmentation generates a single dependent Markov chain of length 40000, where the last 20000 iterations are used for the estimation. The program `Bayes_est.m` returns the posterior mode

```
post_mode =
  0.7468
  0.0962
  0.0582
  0.0430
  0.0557
```

Furthermore, in one run, the command

```
B_mean_PS=PS_stats.B_mean_PS; B_std_PS=PS_stats.B_std_PS; B_CI_PS=PS_stats.B_CI_PS;
[ B_mean_PS B_std_PS B_CI_PS]
```

delivered the following quantities obtained with parameter simulation

```
0.7351 0.0755 0.5815 0.8757
0.0987 0.0292 0.0436 0.1570
0.0610 0.0267 0.0119 0.1156
0.0472 0.0252 0.0047 0.1003
0.0581 0.0272 0.0088 0.1134
```

The first and second column provide posterior means and standard deviations. The third and fourth column contains simulated $2,5\%$ and $97,5\%$ posterior quantiles. E.g., [0.5815, 0.8757] is a 95% Bayes CI for the proportion of individuals having value 1 of the sensitive variable. The above estimates, which were obtained by our MATLAB program, are close to the estimates in Tang et al. (2009), Table 9, page 347. Additionally, our program produces estimates based on multiple imputations and Rao-Blackwellization. In particular, the command

```
B_mean_MI=MI_stats.B_mean_MI; B_std_MI=MI_stats.B_std_MI; B_CI_MI=MI_stats.B_CI_MI;
[B_mean_MI B_std_MI B_CI_MI]
```

returned point estimates `B_mean_MI` (first column), precision measures `B_std_MI` (second column) as well as lower and upper bounds of 95% confidence intervals (third and fourth column):

```
0.7418 0.0735 0.5949 0.8810
0.0972 0.0255 0.0456 0.1443
0.0594 0.0243 0.0127 0.1063
0.0452 0.0232 0.0025 0.0911
0.0564 0.0249 0.0076 0.1038
```

Analogously, the command

```
B_mean_RB=RB_stats.B_mean_RB; B_std_RB=RB_stats.B_std_RB; B_CI_RB=RB_stats.B_CI_RB;
[B_mean_RB B_std_RB B_CI_RB]
```

showed the output

```
0.7350 0.0726 0.5900 0.8725
0.0985 0.0252 0.0475 0.1450
0.0611 0.0240 0.0150 0.1075
0.0472 0.0229 0.0050 0.0925
0.0582 0.0246 0.0100 0.1050
```

Example 2: Version of the two-trial unrelated question model (UQM)

Let us consider the variant of the two-trial UQM from Section 2 in Bourke and Moran (1988) that does not assume independence between the sensitive variable X and the nonsensitive variable Y. According to the data in Bourke and Moran (1988), Table 1, we define

```
p=[0.7 0.3];q=1-p;
C=[
1  p(1)*p(1)  q(1)*q(1)  0;
0  p(1)*q(1)  p(1)*q(1)  0;
0  p(1)*q(1)  p(1)*q(1)  0;
0  q(1)*q(1)  p(1)*p(1)  1;
1  p(2)*p(2)  q(2)*q(2)  0;
0  p(2)*q(2)  p(2)*q(2)  0;
0  p(2)*q(2)  p(2)*q(2)  0;
0  q(2)*q(2)  p(2)*p(2)  1];
nn=[137 271 253 566; 512 291 215 322];
L=[1 1000 501];de=[1 1 1 1]; al=0.05;
[PS_stats, MI_stats, RB_stats, post_mode, Iter]= Bayes_est(nn,C,L,de,al)
```

Notice, `C` is a 8×4 matrix, because this design needs $S = 2$ samples. The MATLAB program `Bayes_est.m` returns especially the posterior mode

```
post_mode =
  0.0000
  0.1240
  0.7788
  0.0972
```

In fact, this posterior mode is equal to the ML estimate, because we have applied the uniform prior distribution, i.e, the Dirichlet distribution with parameter $(1, ..., 1)$. This ML estimate can also be found in Bourke and Moran (1988), Table 2.

If the investigator wants to base parameter simulation, multiple imputation and Rao-Blackwellization on 500 independent Markov chains of length 501, where only the last value of each chain is used for the estimation, he or she must type `L=[500 501 501]` instead of `L=[1 1000 501]`.

Applying the Nonrandomized Diagonal Model to Estimate a Sensitive Distribution in Complex Sample Surveys

Heiko Groenitz

Groenitz, H. (2013): Applying the Nonrandomized Diagonal Model to Estimate a Sensitive Distribution in Complex Sample Surveys. Accepted in: Journal of Statistical Theory and Practice.

A Covariate Nonrandomized Response Model for Multicategorical Sensitive Variables

Heiko Groenitz*

2013

Abstract

The diagonal model (DM) is a recently published nonrandomized response (NRR) survey method to collect data on categorical sensitive characteristics Y^*. Based on DM data, the distribution of Y^* can be estimated. In contrast to randomized response (RR) techniques, NRR schemes avoid the use of a randomization device. Due to this fact, survey complexity and study costs decrease. In this article, we assume that not only Y^*, but also nonsensitive characteristics $X_1^*, ..., X_p^*$ are involved in the survey. Then, the aim of this paper is to provide methods to investigate the dependence of Y^* on $X^* = (X_1^*, ..., X_p^*)$. For instance, the influence of sex and profession on income (recorded in income classes) may be under study. In particular, we describe two estimation procedures: Stratum-wise estimation and LR-DM estimation. Stratum-wise estimation is suitable if only few covariate level appear in the sample. LR-DM estimation is based on a logistic regression model for the relation between Y^* and X^* and requires several techniques for generalized linear models (e.g., Fisher scoring). In simulations, we first investigate the convergence behavior of the Fisher scoring algorithm. Subsequently, we illustrate the connection between efficiency of the LR-DM estimation and the degree of privacy protection. Finally, the efficiency of the LR-DM estimation is compared with the efficiency of the stratum-wise estimation.

1 Introduction

To gather data about sensitive characteristics like income and tax evasion, it is not recommendable to ask directly, because direct questioning provokes answer refusal (i.e., missing values) or untruthful answers. Instead, survey designs that protect the respondents' privacy should be applied, because they can improve the respondents' cooperation. The first privacy-protecting survey method was the randomized response (RR) model by Warner (1965). Today there are several RR procedures which enable the estimation of the distribution of a sensitive characteristic. However, in practice, the investigator is sometimes not only interested in the distribution of the sensitive characteristic, but also in the dependence of the sensitive characteristic on certain covariates. For instance, the influence of age and profession on income might be under study.

The first covariate extension of a RR technique can be found in the book of Maddala (1983), p. 54-56, who proposes a model that enables the analysis of the relation between nonsensitive exogenous variables and a binary sensitive variable.

The paper by Scheers and Dayton (1988) extends the randomized response model by Warner (1965) and the unrelated question (UQM) model (see Greenberg et al. (1969)) with covariates. A survey according to the covariate Warner model proceeds as follows: Consider a sensitive characteristic Y^* with two outcomes, say $Y^* = 1$ and $Y^* = 2$, and an arbitrary respondent. Initially, he or she is asked directly for his or her values of p nonsensitive covariates. Subsequently, he or she draws randomly one of the questions:

$$Q^* = 1: \text{ "Is your value of } Y^* \text{ equal to 1"?} \qquad Q^* = 2: \text{ "Is your value of } Y^* \text{ equal to 2"?} \tag{1}$$

The question might be selected by spinning a spinner for example. The selection occurs hidden and the selected question is not revealed to the interviewer. The respondent replies either "yes" or "no", but the interviewer can not identify the respondent's value of the sensitive characteristic. The authors model the dependence of Y^* on the covariables, for example, by a logistic regression model, and describe methods to maximize the likelihood function. In the case of the UQM, question $Q^* = 2$ would contain a nonsensitive attribute, such as "Are you born in the first quarter of the year?". Within a real data study, the influence

*Philipps-University Marburg, Department for Statistics (Faculty 02), Universitätsstraße 25, 35032 Marburg, Germany (e-mail: groenitz@staff.uni-marburg.de).

of the GPA (grade point average) on academic cheating behavior is investigated. Additional details of this study, especially a comparison between the estimations based on the covariate UQM and an anonymous questionnaire, are available in Scheers and Dayton (1987).

The work by van der Heijden and van Gils (1996) presents a covariate version of the RR method by Kuk (1990). Van den Hout et al. (2007) deal with the analysis of the relation between multiple sensitive characteristics and covariates where the sensitive data are gathered by randomized response methods. They present a real data example regarding social benefit fraud, more precisely the illegal receipt of unemployment benefit in the Netherlands. In particular, the relation between the binary sensitive questions "Is the number of your job applications less then required?" and "Do you conduct any work without reporting this?" and certain covariates (sex, age and an indicator whether the respondent is the main earner in the household) is studied.

In the publications of the previously mentioned authors, RR models are involved in the survey. That means that the respondents have to conduct a random experiment with the help of a randomization device (e.g., spinner or deck of cards). In contrast, nonrandomized response (NRR) techniques, which have been proposed increasingly in the last years, do not need a randomization device. The absence of a randomization device causes a reduction in survey complexity and study costs. Moreover, the respondent would always give the same answer if the survey was conducted again. One such NRR method is the diagonal model (DM) by Groenitz (2012) that is suitable for categorical sensitive characteristics.

After reviewing the DM in Section 2, we consider in Section 3 a survey which includes a sensitive $Y^* \in \{1, ..., k\}$ and nonsensitive characteristics $X_1^*, ..., X_p^*$ where the DM is applied to elicit data about Y^*. Here, the aim of Section 3 is to investigate the influence of $X^* = (X_1^*, ..., X_p^*)$ on Y^*. For this, we present a stratum-wise estimation as well as an estimation that is based on a logistic regression model (LRM). For the latter, extensive material regarding generalized linear models (e.g., Fisher scoring) is required. In Section 4, ample simulations are presented: After a discussion about the convergence behavior of the Fisher scoring algorithm, we analyze the relation between efficiency of the estimation based on a LRM and the degree of privacy protection. Subsequently, we compare the efficiency of the estimation based on a LRM with the efficiency of the stratum-wise estimation.

2 The diagonal model

Groenitz (2012) proposes a nonrandomized response model for multichotomous sensitive variables, namely the diagonal model. This model enables the estimation of the distribution of a sensitive characteristic Y^* with codomain $\{1, ..., k\}$ by the frequencies of certain nonrandomized answers A^*, which depend on an auxiliary variable $W^* \in \{1, ..., k\}$. The auxiliary variable is assumed to be nonsensitive and independent from Y^* with a known distribution $\mathbb{P}_{W^*}$. Moreover, we assume that the interviewer does not know the respondents' values for W^*. Every respondent should give an answer according to

$$A^* := [(W^* - Y^*) \bmod k] + 1. \tag{2}$$

Instead of presenting this formula to the respondents, who may be not familiar with the modular arithmetic, every respondent is given a table where he or she can find the answer to give. For example for $k = 5$, such a table is given by

Y^*/W^*	$W^* = 1$	$W^* = 2$	$W^* = 3$	$W^* = 4$	$W^* = 5$
$Y^* = 1$	1	2	3	4	5
$Y^* = 2$	5	1	2	3	4
$Y^* = 3$	4	5	1	2	3
$Y^* = 4$	3	4	5	1	2
$Y^* = 5$	2	3	4	5	1

Additionally, an example of an answer like "If your value of Y^* equals 3 and your value of W^* equals 1, please give the answer $A^* = 4$" should be included in the questionnaire. The interviewee searches his or her values of Y^* and W^* and gives an answer A^*. Since it is not possible to identify the correct Y^*-value with the help of the answer, we assume that the interviewees cooperate. For instance, W^* could describe the period of birthday of the respondent's mother.

We denote the proportion of units in the population having $Y^* = i$, $W^* = i$ and $A^* = i$ with π_i^*, c_i^* and μ_i^*, respectively. Moreover, let C be the $k \times k$ matrix where every row is a left-cyclic shift of the row

above and the first row is equal to $c^* = (c_1^*, ..., c_k^*)$. The proportions $c_1^*, ..., c_k^*$ are the model parameters and C is referred to as "design matrix of c^*". We have

$$(\mu_1^*, ..., \mu_k^*)^t = C \cdot (\pi_1^*, ..., \pi_k^*)^t. \tag{3}$$

The paper by Groenitz (2012) describes the maximum likelihood (ML) estimation in the case of simple random sampling with replacement, where it turns out that finding an explicit form of the ML estimator is difficult for some samples. However, he shows that the estimation of π^* can be viewed as missing data problem and operated with the expectation maximization (EM) algorithm.

3 Influence of nonsensitive covariates on the sensitive variable

Let us consider a survey involving a categorical, sensitive characteristic $Y^* \in \{1, ..., k\}$ where $k = q + 1$ and a vector of nonsensitive characteristics $X^* = (X_1^*, ..., X_p^*)$. Here, the respondents do not provide their values of Y^*, but give an answer A^* according to the diagonal model. This answer A^* depends on both Y^* and an auxiliary characteristic W^*. We define c^* and the matrix C as in Section 2 and assume throughout the remainder of this article:

- All components of c^* are nonzero (when a c_i^* equaled zero, every answer A^* would restrict the possible Y^*-values).
- The matrix C is invertible.

The aim of this section is to study the dependence of Y^* on X^*. The quantity Y^* is called endogenous characteristic and $X_1^*, ..., X_p^*$ are called exogenous characteristics or covariates or regressors. We consider both deterministic and stochastic covariates.

3.1 The case of deterministic covariates

In this subsection, we assume that the investigator chooses the values of the covariates X^* (i.e., they are fixed and known) and searches persons having the predefined covariate levels. Each person is then requested to give a response A^* according to (2).
For instance, for X_1^*, X_2^*, and Y^* representing sex, profession, and income, respectively, this procedure means that the investigator determines for any combination of sex and profession how many persons possessing this combination are involved into to survey. Then appropriate persons are selected and each person in the sample gives DM answer A^* depending on his or her income and his or her value of the nonsensitive characteristic W^*.

Say n persons are interviewed. Consider for $i = 1, ..., n$ and $j = 1, ..., k$

$$Y_{ij} = \begin{cases} 1, & \text{if person } i \text{ has attribute } Y^* = j \\ 0, & \text{else} \end{cases}, \qquad A_{ij} = \begin{cases} 1, & \text{if person } i \text{ answers } A^* = j \\ 0, & \text{else} \end{cases},$$

let W_i denote the value of W^* corresponding to the i-th person and let x_{ij} represent the value of X_j^* corresponding to the i-th person. Set

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} Y_{11} & \cdots & Y_{1q} \\ \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{nq} \end{pmatrix}, A = \begin{pmatrix} A_1 \\ \vdots \\ A_n \end{pmatrix} = \begin{pmatrix} A_{11} & \cdots & A_{1q} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{nq} \end{pmatrix}, x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

Notice, the realizations of the auxiliary variables W_i and the sensitive variables Y_i are not observed while data on the answers A_i and the regressors x_i are available. We introduce $\pi_{ij} = \mathbb{E}(Y_{ij})$ and $\pi_i = (\pi_{i1}, ..., \pi_{iq})$ as well as $\mu_{ij} = \mathbb{E}(A_{ij})$ and $\mu_i = (\mu_{i1}, ..., \mu_{iq})$. Eventually, we define

$$\pi_j^*(x^*): \text{ proportion of units with } Y^* = j \text{ among the units in the population having } X^* = x^*. \tag{4}$$

In this subsection, we assume throughout

(D1) $Y_1, ..., Y_n$ independent

(D2) $W_1, ..., W_n$ are independent and identically distributed.

(D3) The two quantities $(Y_1^t, ..., Y_n^t)^t$ and $(W_1, ..., W_n)^t$ are independent.

These conditions are fulfilled if (Y^*, X^*) and W^* are independent and if for each covariate level chosen by the investigator, the sample units are drawn by simple random sampling with replacement from the population units having this covariate level where the selection for one covariate level is independent from the selection for the other covariate levels.

Let x^* be one of the covariate levels specified by the investigator, i.e., there is a row of x equal to x^*. The quantity $\pi_j^*(x^*)$ can be estimated from the answers A^* of the persons in the sample having this covariate level x^* according to the estimation procedure in Groenitz (2012) for the diagonal model. Possibly, the EM algorithm must be applied for the estimation.
Let us now assume $g \leq n$ different covariate levels are available. This means that x has g different rows. Then, the set of sample units having the i-th covariate level can be interpreted as stratum i. For this reason, we call the just described estimation method "stratum-wise estimation". One can expect the stratum-wise estimation to be suitable if each stratum contains sufficiently large sample units.

In the sequel, we present an estimation method based on a logistic regression model (LRM). Occasionally, we will call this estimation technique briefly the "LR-DM estimation". LRMs are often applied to analyze the influence of certain covariates on a categorical endogenous characteristic. Some material on LRMs that we need in this article is collected in Appendix A. For the LR-DM estimation, we make the additional assumption:

(D4) There is a $\beta = ({\beta^{(1)}}^t, ..., {\beta^{(q)}}^t)^t$ with $\beta^{(i)} \in \mathbb{R}^{p\times 1}$ so that (Y, x, β) is a logistic regression model.

Of course, the vector β has length $s := pq$ and (D4) includes the independence of $Y_1, ..., Y_n$. Define for $z = (z_1, ..., z_q)$

$$h : z \mapsto (h_1(z), ..., h_q(z)) = \left(\tfrac{e^{z_1}}{1+e^{z_1}+...+e^{z_q}}, \quad \ldots, \tfrac{e^{z_q}}{1+e^{z_1}+...+e^{z_q}}\right), \text{ and } \mathbf{x}_i := \begin{pmatrix} x_i & & \\ & \ddots & \\ & & x_i \end{pmatrix} \in \mathbb{R}^{q\times pq}, \quad (5)$$

and $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$. Then, we have $\pi_i = h((\mathbf{x}_i\beta)^t)$. To estimate β from the LRM (Y, x, β), we have to make a detour via the answers collected in A, because Y is not observed. Let $C(1:q, j) \in \mathbb{R}^q$, $j = 1, ..., q+1$, denote the j-th column of C without the last entry, set $\tilde{c}_j = C(1:q, j) - C(1:q, q+1)$ for $j = 1, ..., q$, and define the $q \times q$ matrix $\tilde{C} := [\tilde{c}_1|\tilde{c}_2|\ldots|\tilde{c}_q]$. We introduce the map

$$m(z) = m(z_1, ..., z_q) = \left[\tilde{C} \cdot \begin{pmatrix} h_1(z) \\ \vdots \\ h_q(z) \end{pmatrix} + C(1:q, k)\right]^t. \quad (6)$$

The following theorem contains an important observation:

Theorem 1 *$(A, x, \beta, \mathbf{x}, m)$ is a generalized linear model (GLM).*

Proof: We must verify that the definition for a GLM (see Appendix B.1) is fulfilled. Since A_i is a function of Y_i and W_i, the independence of $A_1, ..., A_n$ follows. The (discrete) density of A_i is given by

$$f_{A_i}(a_1, ..., a_q) = \mu_{i1}^{a_1} \cdots \mu_{iq}^{a_q} \cdot \mu_{ik}^{1-a_1-...-a_q} \cdot 1_{\mathcal{A}}(a_1, ..., a_q), \quad a_i \in \mathbb{R},$$

where $A = \{(a_1, ..., a_q) : a_i \in \{0, 1\}, a_1 + ... + a_q \leq 1\}$. Set $\Theta = \mathbb{R}^{1\times q}$, $\Psi = \{1\}$ and for $\theta \in \Theta$, $\psi \in \Psi$, $y \in R^{1\times q}$

$$f_{\theta,\psi}(y) = c(y, \psi) \cdot e^{\frac{\theta y^t - b(\theta)}{\psi}} \text{ where } c(y, \psi) = 1_{\mathcal{A}}(y) \text{ and } b(\theta) = log(1 + e^{\theta_1} + ... + e^{\theta_q}).$$

The distribution corresponding to $f_{\theta,\psi}(y)$ is denoted with $\mathbb{P}_{\theta,\psi}$. Consequently, $(\mathbb{P}_{\theta,\psi})_{\theta\in\Theta,\psi\in\Psi}$ is a simple, q-parametric exponential family with scale parameter and we have for $\psi = 1$: For all $i = 1, ..., n$, the distribution of A_i belongs to $(\mathbb{P}_{\theta,\psi})_{\theta\in\Theta}$. Thus, the distribution assumption in Appendix B.1 is satisfied.

The function h is invertible with

$$h^{-1}(w_1, ..., w_q) = (\log \frac{w_1}{w^*}, \ldots, \log \frac{w_q}{w^*}) \text{ where } w^* := 1 - (w_1 + ... + w_q). \quad (7)$$

Applying a chain rule, it suffices to show that the matrix $\tilde{C}$ is regular to ensure the reversibility of m. Assume $\tilde{C}$ is not regular. Then, this matrix has eigenvalue zero, i.e., there is a vector $v = (v_1, ..., v_q)^t \neq 0$

with $\tilde{C}v = 0$. Denoting the $q \times q$ identity matrix by I_q we can write $\tilde{C} = [I_q|(0,...,0)^t] \cdot C \cdot [I_q|(-1,...,-1)^t]^t$. It follows that $0 = [I_q|(0,...,0)^t] \cdot C \cdot (v_1,...,v_q, -\sum_{j=1}^{q} v_j)^t =: [I_q|(0,...,0)^t] \cdot U$. Thus, the first q entries of U are zero. By taking the sum of these q numbers, we can conclude that the k-th entry of U is also zero. Altogether, C has eigenvalue zero. Because we assumed C to be invertible, this is a contradiction. Hence, $\tilde{C}$ is regular.

Finally, we have

$$\mu_i = (\mu_{i1},...,\mu_{iq}) \; = \; m\left((\mathbf{x}_i\beta)^t\right), \tag{8}$$

which completes the proof.

□

Let a_i be an observed realization of A_i. The likelihood function $\beta \mapsto f_{A_1}(a_1) \cdots f_{A_n}(a_n)$ can be maximized via the Fisher scoring algorithm. Some details of this iterative method are provided in Appendix C.1. For our GLM $(A, x, \beta, \mathbf{x}, m)$, we must specify quantities from C.1 as follows. The expectation vector $\mu_i = \mu_i(\beta)$ is given through (8). The Jacobi matrix of m from (6) equals $m'(z) = \tilde{C} \cdot h'(z)$. Here, the Jacobi matrix of h is $h'(z) = [diag(\exp(z) \cdot Q(z)) - \exp(z^t)\exp(z)] \,/\, (Q(z))^2$ with componentwise application of exp and $Q(z) = 1 + e^{z_1} + ... + e^{z_q}$. Furthermore, we have

$$D_i(\beta) = \left[m'((\mathbf{x}_i\beta)^t)\right]^t \text{ and } \Sigma_i(\beta) = Var_\beta(Y_i) = diag(\mu_i(\beta)) - \mu_i(\beta)^t\mu_i(\beta).$$

In GLMs, the asymptotic normality $(F(\hat{\beta}))^{\frac{1}{2}}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, I)$ holds for $n \to \infty$ and $\hat{\beta}$ is approximately $N(\,\beta,\, F^{-1}(\hat{\beta})\,)$-distributed if the total sample size n is sufficiently large (Fahrmeir and Tutz (2010), p. 106). Here, $F(\hat{\beta})$ is the Fisher matrix calculated under $\hat{\beta}$ and $F^{-1}(\hat{\beta})$ can be taken from the last iteration of the Fisher scoring algorithm (cf. Appendix C.1). An estimate for the asymptotical standard error of the i-th component of $\hat{\beta}$ is given by

$$\hat{SE}_{AS}(\hat{\beta}_i) = \sqrt{[F^{-1}(\hat{\beta})]_{ii}}. \tag{9}$$

We now study the estimation of the population parameters $\pi_j^*(x^*)$ from (4). Let us choose a fixed value x^*. Once obtained a maximum likelihood estimate $\hat{\beta}$, we can calculate estimates

$$[\hat{\pi}_1^*(x^*),...,\hat{\pi}_q^*(x^*)] = h((\mathbf{x}^*\hat{\beta})^t), \quad \hat{\pi}_k^*(x^*) = 1 - \hat{\pi}_1^*(x^*) - ... - \hat{\pi}_q^*(x^*), \tag{10}$$

where $\mathbf{x}^*$ is the $q \times s$ design matrix corresponding to x^*. The identity (10) implies that $\hat{\pi}_j^*(x^*)$ is a function of $\hat{\beta}$. In particular, with $H(\beta) = (H_1(\beta),...,H_q(\beta)) = h((\mathbf{x}^*\beta)^t)$ and $H_k(\beta) = h_k((\mathbf{x}^*\beta)^t)$ where $h_k(z) = 1 - h_1(z) - ... - h_q(z)$, we have the equations

$$(\hat{\pi}_1^*(x^*),...,\hat{\pi}_q^*(x^*)) = H(\hat{\beta}) \text{ and } \hat{\pi}_k^*(x^*) = H_k(\hat{\beta}). \tag{11}$$

Using a first-order Taylor approximation of H at β, we obtain

$$\begin{aligned} Var(H(\hat{\beta})) &\approx Var[H(\beta) + J_H(\beta) \cdot (\hat{\beta} - \beta)] = J_H(\beta) \cdot Var(\hat{\beta}) \cdot J_H^t(\beta) \\ &\approx J_H(\hat{\beta}) \cdot \hat{Var}(\hat{\beta}) \cdot J_H^t(\hat{\beta}) = J_h((\mathbf{x}^*\hat{\beta})^t) \cdot \mathbf{x}^* \cdot \hat{Var}(\hat{\beta}) \cdot \mathbf{x}^{*^t} \cdot J_h^t((\mathbf{x}^*\hat{\beta})^t) =: \hat{Var}(H(\hat{\beta})) \end{aligned}$$

where J denotes the Jacobi matrix and $\hat{Var}(\hat{\beta})$ is given by $F^{-1}(\hat{\beta})$. Thus, to estimate the variance of $\hat{\pi}_j^*(x^*)$ $(j = 1,...,q)$, we can use the j-th diagonal element of $\hat{Var}(H(\hat{\beta}))$. Analog, we can derive

$$\hat{Var}(H_k(\hat{\beta})) = J_{h_k}((\mathbf{x}^*\hat{\beta})^t) \cdot \mathbf{x}^* \cdot \hat{Var}(\hat{\beta}) \cdot \mathbf{x}^{*^t} \cdot J_{h_k}^t((\mathbf{x}^*\hat{\beta})^t)$$

with the Jacobi matrix $J_{h_k}(z_1,...,z_k) = (-e^{z_1},...,-e^{z_q})/(Q(z))^2$. The estimated standard errors for the $\hat{\pi}_j^*(x^*)$ are given by taking the square root of the estimated variances for $\hat{\pi}_j^*(x^*)$.

Linear hypotheses concerning β

$$H_0: \quad \mathcal{C}\beta = \rho \quad \text{against} \quad H_1: \quad \mathcal{C}\beta \neq \rho \tag{12}$$

where $\mathcal{C}$ is a $r \times s$ matrix $(r \leq s)$ with full row rank can be tested with the well known Wald statistic (cf. Fahrmeir and Tutz (2010), p. 107)

$$w = (\mathcal{C}\hat{\beta} - \rho)^t \cdot [\mathcal{C} \cdot F^{-1}(\hat{\beta}) \cdot \mathcal{C}^t]^{-1} \cdot (\mathcal{C}\hat{\beta} - \rho),$$

which is asymptotically $\chi^2_{Rank(\mathcal{C})}$-distributed under H_0.

The LR-DM estimation is built on the model structure, especially on (8). To check whether the data fit the relation (8), the Pearson statistic can be applied, provided that we have grouped data such that there is a sufficiently large number of observations in each group. As in Appendix C.1, let $g \leq n$ be the number of different rows of x, i.e., the number of covariate levels, set for $r = 1, ..., g$

$$I_r = \{i \in \{1, ..., n\} : \text{ sample unit } i \text{ possesses covariate level } r\},$$

define n_r to be the number of elements in I_r and assume $i_1 \in I_1$,...,$i_g \in I_g$. The null hypothesis H_0 is given by

$$\mathbb{E}(A_{i_1}) = m((\mathbf{x}_{i_1}\beta)^t), ..., \mathbb{E}(A_{i_g}) = m((\mathbf{x}_{i_g}\beta)^t) \text{ for one } \beta \in \mathbb{R}^s. \tag{13}$$

Set $(\tilde{A}_{r1}, ..., \tilde{A}_{rk}) = n_r^{-1} \sum_{l \in I_r} (A_{l1}, ..., A_{lk})$ and $(\tilde{\mu}_{r1}, ..., \tilde{\mu}_{rq}) = m((\mathbf{x}_{i_r}\hat{\beta})^t)$ and $\tilde{\mu}_{rk} = 1 - \tilde{\mu}_{r1} - ... - \tilde{\mu}_{rq}$. The Pearson statistic P compares $\tilde{A}_{rj}$ and $\tilde{\mu}_{rj}$, in particular, P equals

$$P = \sum_{r=1}^{g} n_r \sum_{j=1}^{k} \frac{(\tilde{A}_{rj} - \tilde{\mu}_{rj})^2}{\tilde{\mu}_{rj}}.$$

If the n_r are sufficiently large, we have approximately $P \sim \chi^2_{(g-p)q}$ under H_0. For more details, see Fahrmeir and Tutz (2010), p. 107. We remark that $\mu_i = m((\mathbf{x}_i\beta)^t) \Leftrightarrow \pi_i = h((\mathbf{x}_i\beta)^t)$. Consequently, the rejection of H_0 from (13) implies that the LRM (Y, x, β) does not fit the observed data.

We provide the self-programmed MATLAB program `fisherscore1.m`, which computes ML estimates for β and $\pi_j^*(x^*)$ (with corresponding standard errors) and assesses the goodness-of-fit, as supplemental material.

3.2 The case of stochastic covariates

In practice, it may occur that the values of the exogenous characteristics are not deterministic (i.e., not determined by the interviewer), but realizations of random variables. For such stochastic regressors, a survey proceeds as follows. Each interviewee is asked directly for his or her values of the nonsensitive covariates $X_1^*, ..., X_p^*$. Afterwards, he or she is requested to give an answer A^* according to the DM answer formula (2).

Let n, Y, A, W_i be defined as in Subsection 3.1, let the random variable X_{ij} represent the value of X_j^* corresponding to the i-th person in the sample and set $X_i = (X_{i1}, ..., X_{ip})$ as well as $X = (X_1^t, ..., X_n^t)^t$.

In this subsection, we have to incorporate the stochastic character of X into our assumptions. In particular, we assume throughout this subsection

(S1) $(Y_1, X_1), ..., (Y_n, X_n)$ are n iid vectors.

(S2) $W_1, ..., W_n$ are iid.

(S3) The two quantities $\begin{pmatrix} Y_1, X_1 \\ \vdots \\ Y_n, X_n \end{pmatrix}$ and $\begin{pmatrix} W_1 \\ \vdots \\ W_n \end{pmatrix}$ are independent.

These requirements are satisfied when (Y^*, X^*) and W^* are independent and the interviewees are selected by simple random sampling with replacement from the population.

Stratum-wise estimation can be conducted analog to Subsection 3.1. To convey the LR-DM estimation as presented in the previous subsection to the case of stochastic regressors, we further assume

(S4) There is a $\beta = (\beta^{(1)^t}, ..., \beta^{(q)^t})^t$ with $\beta^{(i)} \in \mathbb{R}^{p \times 1}$ so that (Y, X, β) is a LRM with stochastic covariates (see Appendix A.2).

We have that $(A_1, X_1), ..., (A_n, X_n)$ are n iid vectors and that $A_1, ..., A_n$ are independent given $X_1 = x_1, ..., X_n = x_n$ (for all values $x_1, ..., x_n$). Moreover, with $\mathbf{X}_i := \begin{pmatrix} X_i & & \\ & \ddots & \\ & & X_i \end{pmatrix} \in \mathbb{R}^{q \times pq}$ and $\mathbf{X} =$

$(\mathbf{X}_1, ..., \mathbf{X}_n)$ as well as m from (6), we have $\mathbb{E}(A_i|X) = m((\mathbf{X}_i\beta)^t)$ and $(A, X, \beta, \mathbf{X}, m)$ is a GLM with stochastic covariates (cf. Appendix B.2).

The maximum likelihood estimation for $\beta \in \mathbb{R}^{s\times 1}$ with $s = pq$ in this GLM with stochastic covariates can be traced back to the ML estimation in a GLM with deterministic covariates (see Appendix C.2). Thus, our program `fisherscore1.m` can also be applied to calculate MLEs in the case of stochastic covariates. The asymptotic normality $(F(\hat{\beta}))^{\frac{1}{2}}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, I)$ of the MLE $\hat{\beta}$ also holds for GLMs with stochastic covariates (Fahrmeir and Tutz (2010), p. 106). Thus, $\hat{\beta}$ has the approximative distribution $N(\,\beta,\, F^{-1}(\hat{\beta})\,)$ when n is sufficiently large. Consequently, an estimate for the asymptotical standard error of $\hat{\beta}_i$ is $\sqrt{[F^{-1}(\hat{\beta})]_{ii}}$. Linear hypotheses (12) can be tested with the Wald statistic (Fahrmeir and Tutz, p.107)

$$W = (\mathcal{C}\hat{\beta} - \rho)^t \cdot [\mathcal{C} \cdot F^{-1}(\hat{\beta}) \cdot \mathcal{C}^t]^{-1} \cdot (\mathcal{C}\hat{\beta} - \rho),$$

which is also in the case of stochastic covariates asymptotically $\chi^2_{Rank(\mathcal{C})}$-distributed under the null hypothesis.

For a fixed covariate level x^*, the population parameters $\pi_j^*(x^*)$ from (4) can be estimated totally analog to Subsection 3.1 by (11). The estimated standard errors for this estimation can be obtained again as in Subsection 3.1.

For grouped data with a sufficiently large number of observations in each group, the goodness-of-fit can be assessed by the Pearson statistics P as in Subsection 3.1, where we have the approximative conditional distribution $P|X = x \quad \sim \quad \chi^2_{(g-p)q}$ under H_0.

4 Simulations

4.1 Convergence behavior of the scoring algorithm

The maximum likelihood estimation for a GLM according to Section 3 requires the maximization of

$$\beta \mapsto \sum_{i=1}^{n} (a_{i1}, ..., a_{ik}) \cdot \log\left(C \cdot (\pi_{i1}, ..., \pi_{ik})^t\right) \tag{14}$$

where a_{ij} is a realization of A_{ij}, π_{ij} depends on β, and log is applied componentwise. It may occur that the function (14) does not possess a maximum. A discussion about the existence of an MLE in general GLMs including further references can be found in Fahrmeir and Tutz (2010), p. 43. Nevertheless, the mathematical conditions for the existence are usually difficult to check in practice. We will illustrate the non-existence with some examples:

1. We first give an example for which we can show by simple analytic methods that a MLE does not exist. Let $Y^* \in \{1, ..., k\}$ (with $\pi_i^* > 0$) be a sensitive variable and assume we have conducted a survey due to the non-covariate diagonal model with n interviewees drawn by a simple random sample with replacement. Define Y, A, W_i as in Subsection 3.1. For observed values a_{ij} of A_{ij}, the log-likelihood is given by

$$\tilde{l}((\pi_1, ..., \pi_k)^t) = \left(\sum_{i=1}^{n} (a_{i1}, ..., a_{ik})\right) \cdot \log\left(C \cdot (\pi_1, ..., \pi_k)^t\right).$$

Set $x = (1, ..., 1)^t \in \mathbb{R}^n$, $\mathbf{X}_i = I_q$, $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, $\beta = h^{-1}(\pi_1^*, ..., \pi_q^*)$ with the link function h^{-1} from (7). With the map m from (6), it follows that $(A, x, \beta, \mathbf{x}, m)$ is a GLM with log-likelihood function

$$l(\beta) = \left(\sum_{i=1}^{n} (a_{i1}, ..., a_{ik})\right) \cdot \log\left(C \cdot H(\beta)\right),$$

where H is a function $\mathbb{R}^{q\times 1} \to \{(x_1, ..., x_k)^t : x_i \in (0, 1), \sum_{i=1}^{k} x_i = 1\}$ with $H(\beta) = (h_1(\beta^t), ..., h_q(\beta^t), 1 - h_1(\beta^t) - ... - h_q(\beta^t))^t$.

Let us now specify $k = 2$, $c^* = (0.6, 0.4)$ and let the number of respondents who give answer 1 and 2 equal 15 and 5, respectively. Suppose that l possesses on $\mathbb{R}$ a maximum $\hat{\beta}$. Then, $H(\hat{\beta})$ would be the maximum of $\tilde{l}$ on the set $\{(x_1, x_2)^t : x_i \in (0, 1), x_1 + x_2 = 1\}$. However, we can easily show that $\tilde{l}$ does not possess

a maximum on $\{(x_1, x_2)^t : x_i \in (0,1), x_1 + x_2 = 1\}$ for above specifications. Due to this contradiction, l has no maximum on $\mathbb{R}$.

2. Let us consider a sensitive Y^* with range $\{1, 2\}$ and exogenous characteristics $X^* = (X_1^*, X_2^*)$ where X_1^* is constant equal to one and $X_2^* \in \{1, 2, 3\}$. We assume stochastic covariates and make the following specifications taken from an example in Scheers and Dayton (1988), Section 3:

$$n = 200, \quad w = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} 0.1587 \\ 0.6826 \\ 0.1587 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} -3.118 \\ 1.218 \end{pmatrix}$$

where w_i is defined to be the proportion of individuals in the universe having attribute $X_2^* = i$. Furthermore, we set $c^* = (0.7, 0.3)$. As before c^* describes the distribution of an auxiliary variable. We have simulated 1000 samples where realizations of A and X are available for each sample. To obtain one sample it suffices to generate absolute frequencies of the covariate levels $(n_1,\ n_2,\ n_3) \sim Multinomial(n, w)$ and to subsequently draw the frequencies of the answers $A^* = j$ for each covariate level from the multinomial distribution with number of trials equal to n_i and cell probabilities $(m(\beta_1 + i\beta_2), 1 - m(\beta_1 + i\beta_2))$.

For each sample, we tried to compute a MLE $\hat{\beta}$ with the self-programed MATLAB program `fisherscore1` and also with the function `glmfit` which is already available in MATLAB. A valid estimate is obtained for most samples, but for some samples the estimation fails. For instance, no problems occur for

covariate level (X_1^*, X_2^*)	$(1,\ 1)$	$(1, 2)$	$(1, 3)$
observations	32	137	31
frequency of $A^* = 1$	16	55	16

(15)

where $\hat{\beta} = \begin{pmatrix} -0.8750 & 0.0999 \end{pmatrix}^t$. Otherwise, the sample with

covariate level (X_1^*, X_2^*)	$(1,\ 1)$	$(1, 2)$	$(1, 3)$
observations	30	144	26
frequency of $A^* = 1$	8	68	18

(16)

leads to $\hat{\beta} = (NaN,\ NaN)^t$ in `fisherscore1` respectively to a complex-valued $\hat{\beta} = (-8.2030+6.2832\mathtt{i},\ 3.9660-3.1416\mathtt{i})^t$ using `glmfit`. The contour plots in Figures 1 and 2 give an illustration of the log-likelihood function for (15) and (16). According to our simulation, non-convergence occured in 5.4% (`fisherscore1`) respectively 7.3% (`glmfit`) of the samples. The difference may be explained by the fact that `fisherscore1` has used serveral starting values whereas user-defined starting values cannot be inputted in `glmfit`.

4.2 Efficiency of LR-DM estimation and degree of privacy protection (DPP)

For the non-covariate diagonal model, Groenitz (2012), Sections 3.5 and 4.2, has shown how the distribution $c^* = (c_1^*, ..., c_k^*)$ of the auxiliary characteristic W^* influences the DPP and efficiency. The goal of this section is to illustrate the influence of c^* for the LR-DM estimation within a simulation study. Here, we consider $k = 4$, $n = 300$, $X^* = (X_1^*, X_2^*)$, where X_1^* is a constant equal to one and X_2^* has codomain $\{1, ..., 5\}$, as well as $\beta = (3.5,\ -1.25,\ 2.5,\ -0.5,\ 2,\ -0.25)^t$ and $w = (1,\ 2,\ 3,\ 2,\ 1)/9$. The i-th component of w denotes the proportion of people in the population with level $X_2^* = i$. The entry (i, j) of the matrix

$$\begin{pmatrix} 0.4015 & 0.3127 & 0.2435 & 0.0423 \\ 0.2143 & 0.3534 & 0.3534 & 0.0789 \\ 0.0975 & 0.3403 & 0.4370 & 0.1252 \\ 0.0399 & 0.2949 & 0.4863 & 0.1789 \\ 0.0153 & 0.2392 & 0.5063 & 0.2392 \end{pmatrix} \tag{17}$$

denotes the proportion of units with $Y^* = j$ among the units in the universe having covariate value $X_2^* = i$. That is, the matrix entries equal the $\pi_j^*(x^*)$ according to (4). Imagine that Y^* describes income classes where category $Y^* = 1$ ($Y^* = k$) represents low (high) income, and covariable X_2^* describes age classes where $X_2^* = 1$ ($X_2^* = 5$) indicates a low (high) age. Then, (17) might be realistic relative frequencies, because income often grows with increasing age.

We can measure the efficiency of estimators $\hat{\pi}_j^*(x^*)$ for each covariate level x^* (for our setup, we have $x^* \in \{(1,\ i) : i = 1, ..., 5\}$) by

$$trace\left[MSE\left(\hat{\pi}_1^*(x^*), ..., \hat{\pi}_k^*(x^*)\right)\right] = MSE(\hat{\pi}_1^*(x^*)) + ... + MSE(\hat{\pi}_k^*(x^*)). \tag{18}$$

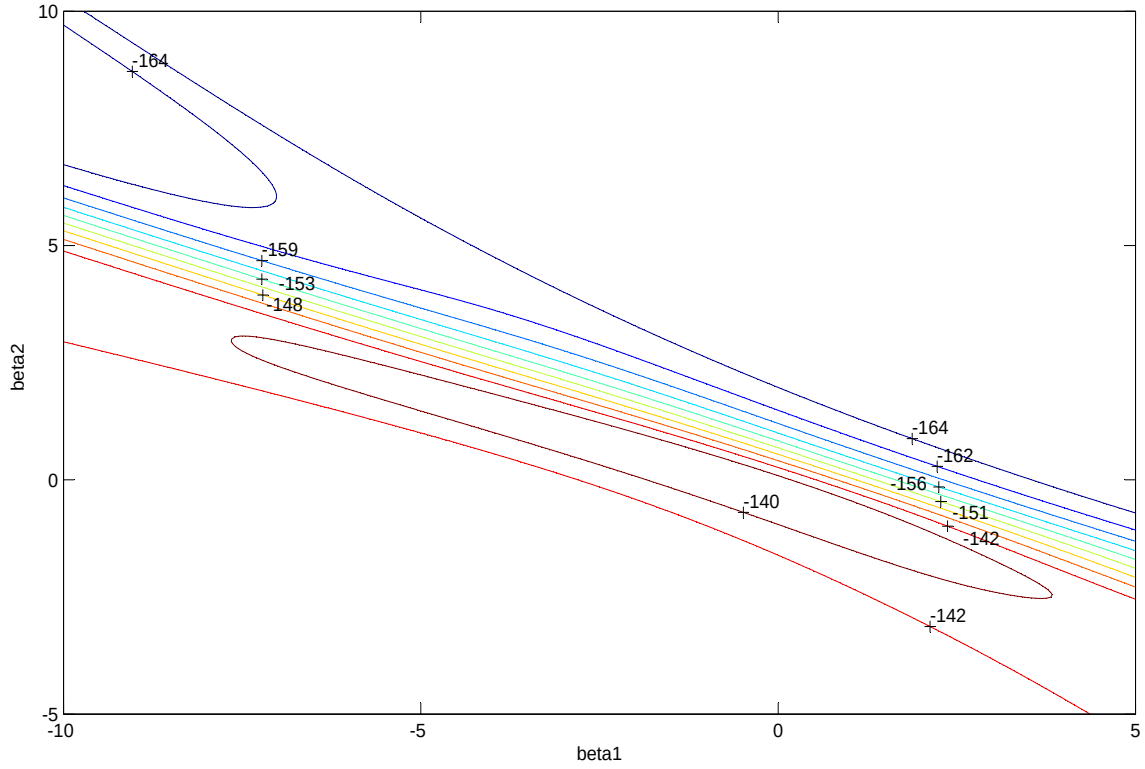


Figure 1: Contour plot for the log-likelihood l corresponding to (15): Consider the isoline $l = -140$. Outside this isoline we have $l < -140$ and the maximum of l is located in the domain $\{\beta : l(\beta) \geq -140\}$. In particular, the maximum is $(-0.875, 0.0999)$ with likelihood value -136.93.

In our simulations, we consider several vectors c^*. As in Groenitz (2012), we use the standard deviation of the vector c^*, denoted by $\sigma = std(c^*) \in [0, \sqrt{1/k}]$, to quantify the DPP. In other words, we measure the closeness of the distribution of W^* to a degenerate and a uniform distribution.

The simulations start with the draw of 500 vectors $c^* = (c_1^*, ..., c_4^*)$ which are uniformly scattered on $\{(x_1, ..., x_4) \in [0,1]^4 : x_1 + ... + x_4 = 1\}$. One such c^* can be generated as follows: Simulate (c_1^*, c_2^*, c_3^*) from a Dirichlet distribution with parameter $(1, 1, 1, 1)$, see Gentle (1998), p. 111, and define $c_4 = 1-(c_1+c_2+c_3)$.

For each drawn c^*, we compute the standard deviation of c^* as measure for the DPP and generate 100 samples. To obtain one sample, we draw $(n_1, ..., n_5) \sim Multinomial(n, w)$. This implies that we have stochastic covariates. Afterwards, we draw the frequencies of the responses $A^* = j$ for each covariate level x^* from the multinomial distribution with parameters n_i and

$$\left[m_1((\mathbf{x}^*\beta)^t), \quad ... \quad , m_q((\mathbf{x}^*\beta)^t), \quad 1 - \sum_{j=1}^{q} m_j((\mathbf{x}^*\beta)^t) \right]. \tag{19}$$

As before, $\mathbf{x}^*$ denotes the $q \times s$ design matrix corresponding to x^*. As already mentioned in Section 4.1, the ML estimation for β may fail. We delete all samples in which `fisherscore1` does not converge. For each of the remaining samples, we calculate $\hat{\pi}_j^*(x^*)$ from $\hat{\beta}$, see (10). Based on the realizations of $\hat{\pi}_j^*(x^*)$, we calculate the empirical MSE. That is, we compute an estimate $\hat{\mathbb{E}}((\hat{\pi}_j^*(x^*) - \pi_j^*(x^*))^2 \,|\, \mathcal{B})$ with the event $\mathcal{B} = \{$ MLE exists $\}$. The quantity (18) is then estimated by the simulated MSE sum $\sum_{j=1}^{4} \hat{\mathbb{E}}((\hat{\pi}_j^*(x^*) - \pi_j^*(x^*))^2 \,|\, \mathcal{B})$.

As soon as the simulations for the randomly drawn c^* have been completed, we repeat the procedure with the vectors $c^{*(1)}, ..., c^{*(6)} \in \mathbb{R}^4$ according to Theorem 2b in Groenitz (2012) for the corresponding degrees of privacy protection $\sigma_i = i/12$. Clearly, the σ_i $(i = 1, ..., 6)$ are equidistant points in the range of the standard deviation.

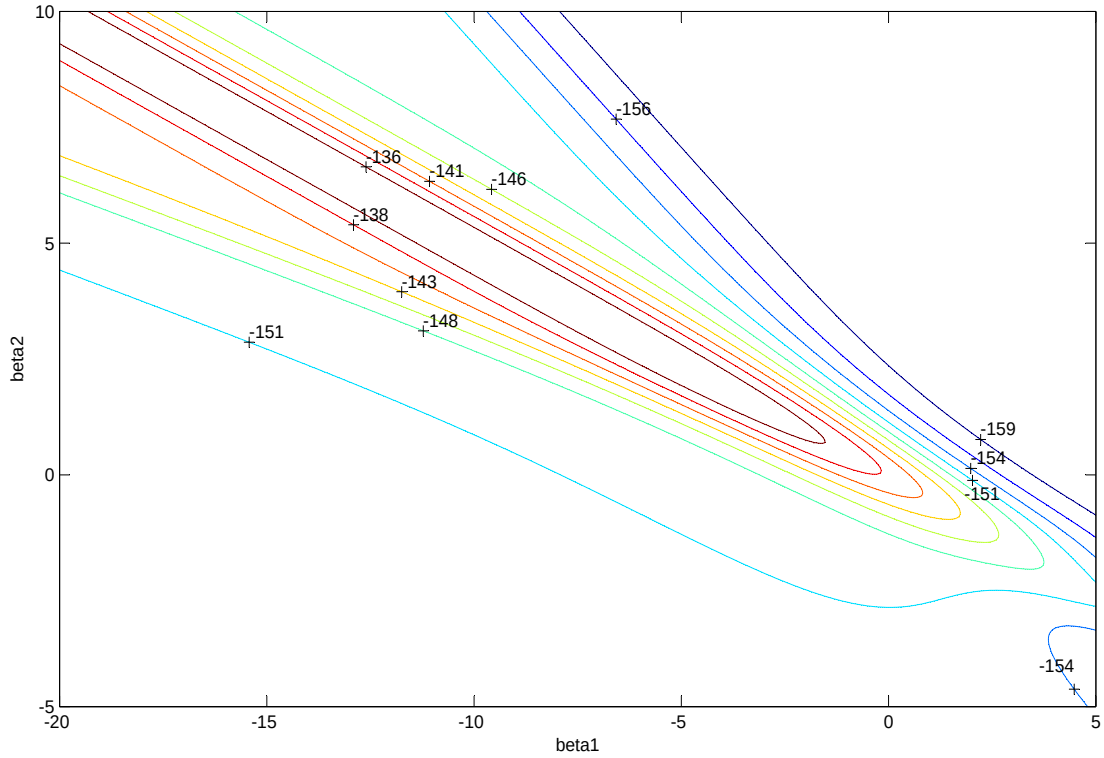


Figure 2: Contour plot for the log-likelihood l corresponding to (16): If we are located in $(0,0)^t$ and move upwards and to the left in the picture, we will successively find vectors β with increasing likelihood.

Due to Figure 3, the nonconvergence probability seems to have a lower bound that depends on σ. The nonconvergences rates of $c^{*(i)}$ decrease from $c^{*(1)}$ to $c^{*(6)}$ and are close to this lower bound. However, $c^{*(1)}$ and $c^{*(2)}$ are impractical, because the ML estimation often fails. Let us now consider Figure 4. For any covariate level, the point cloud for the randomly drawn vectors has a lower bound. The crosses ($\times$) for $c^{*(2)}, ..., c^{*(6)}$ ($c^{*(1)}$ was omitted due to the high nonconvergence rate) are located quite accurate on this bound. Thus, we conclude that the $c^{*(i)}$ are efficient choices for $\mathbb{P}_{W^*}$ for the corresponding degrees of privacy protection. If we connect the 5 crosses, we obtain a strictly monotonically decreasing polygonal curve. That means a larger degree of privacy protection is associated with smaller efficiency. Altogether, the observed influence of $\mathbb{P}_{W^*}$ on efficiency of the LR-DM estimation coincides with the results for the non-covariate diagonal model.

Hence, the interviewer should fix a medium value of σ and determine the vector c^* via Theorem 2b from Groenitz (2012). Finally, an auxiliary attribute W^* should adapted on the chosen c^*.

4.3 Efficiency comparison

Let us consider a sensitive characteristic $Y^* \in \{1, ..., k\}$ and covariates $X^* = (X_1^*, X_2^*)$ where X_1^* is constant equal to one and X_2^* is nonsensitive and can attain the outcomes $1, ..., g^*$. We specify $k = 3$, $c^* = (2/3, 1/6, 1/6)$, and $g^* \in \{3, 5\}$, i.e., either three or five covariate levels appear in the population. Moreover, we assume that the relation between Y^* and X^* follows a logistic regression model with $\beta = (3.50, -1.25, 2.50, -0.50)^t$. We have the following proportions of units having $Y^* = j$ among the units in the population with covariate level x^*:

$g^* = 3$ covariate levels

x^* / j	1	2	3
(1,1)	0.5307	0.4133	0.0559
(1,2)	0.3315	0.5465	0.1220
(1,3)	0.1732	0.6045	0.2224

$g^* = 5$ covariate levels

x^* / j	1	2	3
(1,1)	0.5307	0.4133	0.0559
(1,2)	0.3315	0.5465	0.1220
(1,3)	0.1732	0.6045	0.2224
(1,4)	0.0777	0.5741	0.3482
(1,5)	0.0310	0.4845	0.4845

(20)

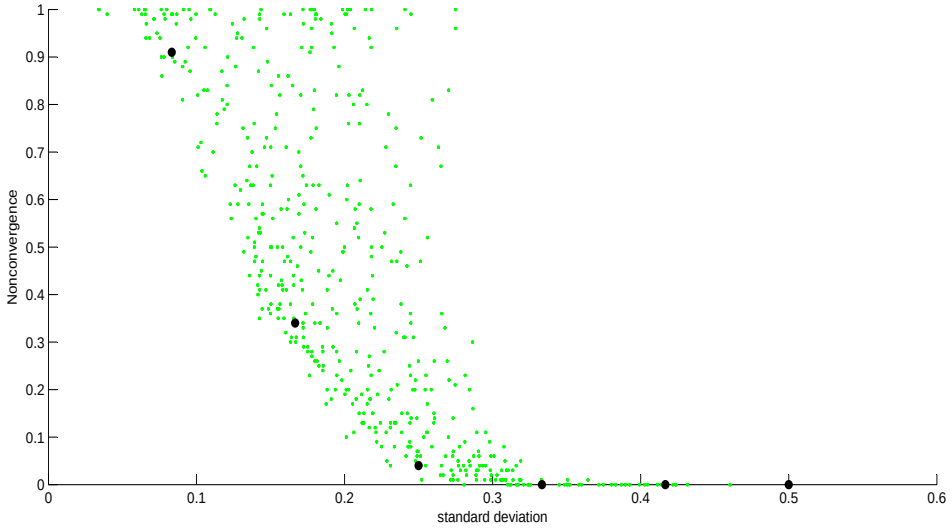


Figure 3: Nonconvergence rates in dependence of the standard deviation σ. A small point corresponds to a vector c^* that is drawn randomly. The boldfaced black dots belong to the $c^{*(i)}$.

Similar to Section 4.2 the proportions in (20) might be realistic proportions for Y^* and X_2^* describing income and age classes, respectively. Notice, the elements of the tabulars in (20) equal the $\pi_j^*(x^*)$ according to (4). We consider sample sizes $n \in \{100, 200, 300, 400\}$ and several specifications for w where the i-th component of w denotes the relative frequency of units in the universe having $x^* = (1, i)$:

$$\begin{aligned} g^* = 3: &\quad w^{(1)} = (1,\,1,\,1)/3 \text{ and } w^{(2)} = (1,\,2,\,3)/6 \\ g^* = 5: &\quad w^{(1)} = (1,\,1,\,1,\,1,\,1)/5 \text{ and } w^{(2)} = (1,\,2,\,3,\,2,\,1)/9 \end{aligned} \tag{21}$$

The aim of this subsection is to compare the efficiency of two estimation procedures: On the one hand, we estimate $\pi_j^*(x^*)$ from (20) according to the LR-DM estimation. On the other hand, a stratum-wise estimation is conducted.

For each specification of (g^*, w, n), we simulate 1000 samples. Each sample consists of $n_i = round(w_i \cdot n)$ interviewees with covariate level $x^* = (1, i)$. Here, the operator *round* means rounding to the nearest integer and w_i is the i-th component of w. This situation corresponds to deterministic covariates. For covariate level $x^* = (1, i)$, we draw the frequencies of the replies A^* from a multinomial distribution analog to the description around (19). Since the ML estimation for β may fail, we delete all samples in which `fisherscore1` does not converge. For each of the remaining samples, we calculate estimates for $\pi_j^*(x^*)$ - once by LR-DM estimation and once by stratum-wise estimation.

For each considered estimator, we compute the average and the empirical mean squared error (MSE) from the available realizations. This means that we obtain estimates for expectation and MSE of the estimators. An excerpt of the simulation output can be found in the Tables 1 and 2.

We first regard five covariate levels. It turns out, that the nonconvergence rates decrease strongly with increasing sample size (for $w^{(1)}$: reduction from 19,6% ($n = 100$) to 0,3% ($n = 400$); for $w^{(2)}$: reduction from 13% ($n = 100$) to 0,2% ($n = 400$)). This coincides with the theoretic result that the existence of a MLE for β in GLMs is asymptotically guaranteed (cf. Fahrmeir and Tutz (2010), p.44).

Let us now focus on the estimation of the conditional proportions $\pi_j^*(x^*)$. On average, the estimates calculated according to both LR-DM and stratum-wise estimation are close to the true values of $\pi_j^*(x^*)$. Regarding efficiency, the empirical MSEs of the estimates decreases if the sample size grows. Moreover, the empirical MSEs corresponding to LR-DM estimation are always smaller than the MSEs corresponding to stratum-wise estimation. The quotient of empirical MSE for LR-DM estimation divided by empirical MSE for stratum-wise estimation attains values between 17% and 93% where it is mostly less than 60%. That is, the estimation precision can be improved significantly by using the functional form (22) from Appendix A.1.

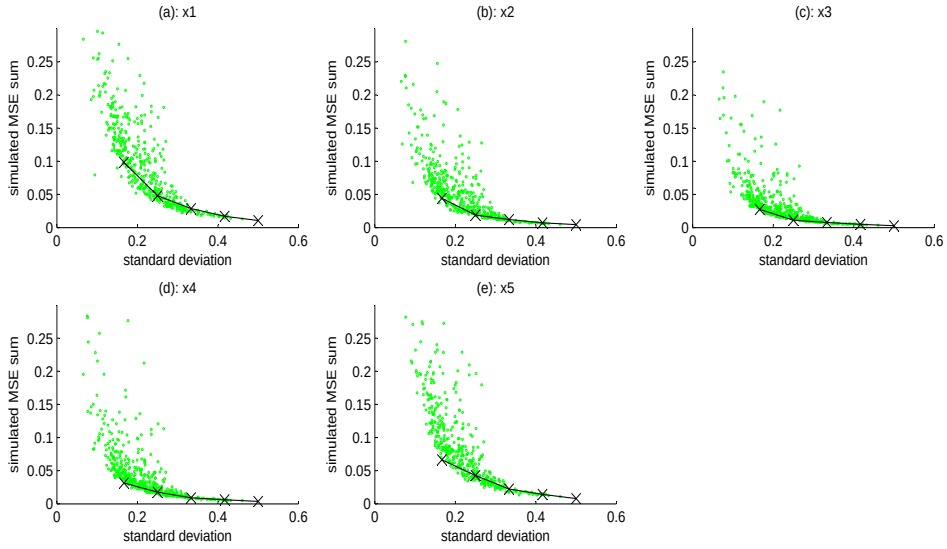


Figure 4: Plots of the simulated MSE sum against the standard deviation for each covariate level. E.g., `x1` in the heading of the first plot, means the covariate level $x^* = (1, 1)$. A small point corresponds to a vector c^* that is drawn randomly. The boldfaced black dots belong to $c^{*(2)}, ..., c^{*(6)}$.

LR-DM estimation average of the estimates				
	covariate level	$Y^* = 1$	$Y^* = 2$	$Y^* = 3$
$n = 300$, $w^{(1)}$	$(1, 1)$	0.5411	0.3972	0.0617
	$(1, 2)$	0.3283	0.5478	0.1239
	$(1, 3)$	0.1718	0.6127	0.2156
	$(1, 4)$	0.0835	0.5774	0.3390
	$(1, 5)$	0.0385	0.4757	0.4858
	non-conv	3		
$n = 300$, $w^{(2)}$	$(1, 1)$	0.5289	0.4081	0.0629
	$(1, 2)$	0.3247	0.5509	0.1244
	$(1, 3)$	0.1712	0.6111	0.2176
	$(1, 4)$	0.0851	0.5715	0.3434
	$(1, 5)$	0.0414	0.4731	0.4855
	non-conv	4		

Stratum-wise estimation average of the estimates				
	covariate level	$Y^* = 1$	$Y^* = 2$	$Y^* = 3$
$n = 300$, $w^{(1)}$	$(1, 1)$	0.5295	0.3992	0.0713
	$(1, 2)$	0.3322	0.5419	0.1259
	$(1, 3)$	0.1738	0.5971	0.2292
	$(1, 4)$	0.0918	0.5738	0.3344
	$(1, 5)$	0.0569	0.4710	0.4721
$n = 300$, $w^{(2)}$	$(1, 1)$	0.5052	0.4076	0.0872
	$(1, 2)$	0.3288	0.5443	0.1269
	$(1, 3)$	0.1739	0.6067	0.2194
	$(1, 4)$	0.0931	0.5642	0.3427
	$(1, 5)$	0.0694	0.4665	0.4641

Table 1: The left (right) part of the table contains the averages of the estimates for $\pi_j^*(x^*)$ according to the LR-DM estimation (stratum-wise estimation). The entry "non-conv" counts how often `fisherscore1` did not converge.

The aforementioned observations for five covariate levels can be also found in the case of three covariate levels. The only noticeable difference is that higher nonconvergence rates of `fisherscore1` occur in the three level case. Altogether, we conclude the major result of this section: If the logistic regression model fits the data, the use of the functional structure (22) leads to a considerably reduction of the MSE.

5 Summary

In this article, we have considered a survey with a sensitive attribute $Y^* \in \{1, ..., k\}$ and nonsensitive characteristics $X^* = (X_1^*, ..., X_p^*)$ where the collection of data on Y^* is conducted with the nonrandomized diagonal model. To examine the dependence of Y^* on X^*, we have introduced the stratum-wise estimation and the LR-DM estimation, which is built on a logistic regression model for the relation between Y^* and X^*. For the LR-DM estimation, maximum likelihood estimates must be computed iteratively where the Fisher scoring algorithm is helpful. In simulations, we investigated the convergence probabilities of Fisher scoring and discussed how the efficiency of the LR-DM estimation depends on the degree of privacy protection. In a further part of the simulation study, we considered a situation where the data fit a logistic regression model. We found out that the application of the functional relation between the proportion of units in the population having outcome $Y^* = j$ and the covariates leads to considerably smaller mean squared errors than a stratum-wise estimation.

LR-DM estimation (MSEs)				
	covariate level	$Y^*=1$	$Y^*=2$	$Y^*=3$
$n=300$, $w^{(1)}$	$(1,1)$	0.0129	0.0102	0.0022
	$(1,2)$	0.0063	0.0057	0.0039
	$(1,3)$	0.0048	0.0059	0.0051
	$(1,4)$	0.0029	0.0053	0.0054
	$(1,5)$	0.0014	0.0103	0.0112
	non-conv.	3		
$n=300$, $w^{(2)}$	$(1,1)$	0.0178	0.0142	0.0024
	$(1,2)$	0.0066	0.0062	0.0035
	$(1,3)$	0.0044	0.0052	0.0038
	$(1,4)$	0.0029	0.0056	0.0056
	$(1,5)$	0.0017	0.0142	0.0159
	non-conv.	4		

Stratum-wise estimation (MSEs)				
	covariate level	$Y^*=1$	$Y^*=2$	$Y^*=3$
$n=300$, $w^{(1)}$	$(1,1)$	0.0147	0.0142	0.0065
	$(1,2)$	0.0146	0.0149	0.0094
	$(1,3)$	0.0117	0.0169	0.0132
	$(1,4)$	0.0077	0.0152	0.0143
	$(1,5)$	0.0058	0.0146	0.0147
$n=300$, $w^{(2)}$	$(1,1)$	0.0260	0.0242	0.0112
	$(1,2)$	0.0129	0.0144	0.0084
	$(1,3)$	0.0079	0.0104	0.0081
	$(1,4)$	0.0070	0.0133	0.0129
	$(1,5)$	0.0102	0.0264	0.0252

Table 2: Empirical mean squared errors (MSEs) of the estimates for $\pi_j^*(x^*)$ using the LR-DM procedure and the stratum-wise estimation.

Appendix

For the LR-DM estimation we need some material regarding logistic regression models (LRMs) and generalized linear models (GLMs). Although LRMs and GLMs are well-known (e.g., Fahrmeir and Tutz (2010)), we briefly mention some facts in this appendix to increase the readability of the paper.

A Logistic regression models (LRMs)

A.1 LRMs with deterministic covariates

Consider random variables Y_{ij} $(i=1,...,n;\ j=1,...,q)$, define the random vectors $Y_i=(Y_{i1},...,Y_{iq})$ and the random matrix $Y=(Y_1^t,...,Y_n^t)^t$. Let x_{ij} $(i=1,...,n; j=1,...,p)$ be real numbers, define $x_i=(x_{i1},...,x_{ip})$ and the deterministic matrix $x=(x_1^t,...,x_n^t)^t$. Moreover, assume $\beta^{(1)},...,\beta^{(q)}\in\mathbb{R}^{p\times 1}$ and set $\beta=({\beta^{(1)}}^t,...,{\beta^{(q)}}^t)^t$. The triple (Y,x,β) is called logistic regression model, if

1. $Y_1,...,Y_n$ are independent and the random vector $(Y_{i1},...,Y_{iq},1-\sum_{j=1}^q Y_{ij})$ is multinomially distributed with number of trials equal to one.

2. The equations
$$\mathbb{P}(Y_{ij}=1)=\frac{e^{x_i\beta^{(j)}}}{1+e^{x_i\beta^{(1)}}+...+e^{x_i\beta^{(q)}}}\quad (i=1,...,n; j=1,...,q) \tag{22}$$
hold for the cell probabilities.

When (Y,x,β) is a LRM, we set $k=q+1$, $Y_{ik}=1-\sum_{j=1}^q Y_{ij}$ and can conclude that

$$\mathbb{P}(Y_{ij}=1)\,/\,\mathbb{P}(Y_{ik}=1)=e^{x_i\beta^{(j)}}\quad (j=1,...,q). \tag{23}$$

In applications, LRMs are useful to study the dependence of a categorical characteristic $Y^*\in\{1,...,k\}$ with $k=q+1$ on a vector of covariates $X^*=(X_1^*,...,X_p^*)$. Here, one considers a sample of size n and the Y_{ij} are given by

$$Y_{ij}=1\ \ (Y_{ij}=0) \text{ if sample unit } i \text{ possesses outcome } Y^*=j\ \ (Y^*\neq j),$$

whereas the value of X^* corresponding to the i-th sample unit is denoted with x_i. According to (23), the components of the parameter β can be interpreted in the following way: E.g., an increase by 1 in the second covariate causes a change in the odds ratio $\mathbb{P}(Y_{ij}=1)\,/\,\mathbb{P}(Y_{ik}=1)$ by the factor $e^{\beta_2^{(j)}}$.

A.2 LRMs with stochastic covariates

In practice, the values of the covariates are often not deterministic, but realizations of random quantities. This motivates to consider also LRMs with stochastic regressors. Define Y and β as in A.1, let X_{ij} $(i=1,...,n; j=1,...,p)$ be random variables, set $X_i=(X_{i1},...,X_{ip})$ and $X=(X_1^t,...,X_n^t)^t$. The triple (Y,X,β) is called a LRM with stochastic covariates, if the following properties are satisfied for every value x of X:

1. The $Y_1,...,Y_n$ are independent given $X=x$ and the conditional distribution of the vector $(Y_{i1},...,Y_{iq},1-\sum_{j=1}^q Y_{ij})$ given $X=x$ is a multinomial distribution with number of trials equal to one.

2. The identities

$$\mathbb{P}(Y_{ij} = 1 \,|\, X = x) = \frac{e^{x_i\beta^{(j)}}}{1 + e^{x_i\beta^{(1)}} + ... + e^{x_i\beta^{(q)}}} \quad (i = 1, ..., n; j = 1, ..., q) \tag{24}$$

hold (x_i is the i-th row of x).

B Generalized linear models (GLMs)

As preparatory work, we need the following definition: A family $(\mathbb{P}_{\theta,\psi})_{\theta\in\Theta,\psi\in\Psi}$ of distributions on the Borel σ-algebra over $\mathbb{R}^q$ is called "simple, q-parametric exponential family with scale parameter" if functions $c : \mathbb{R}^q \times \Psi \to [0, \infty)$ and $b : \Theta \to \mathbb{R}$ exist with the property: Any $\mathbb{P}_{\theta,\psi}$ has a density of the form

$$f_{\theta,\psi}(y) = f_{\theta,\psi}(y_1, ..., y_q) = c(y, \psi) \cdot e^{\frac{\theta y^t - b(\theta)}{\psi}} \quad (y \in \mathbb{R}^q).$$

B.1 GLMs with deterministic covariates

Consider random variables Y_{ij} $(i = 1, ..., n;\ j = 1, ..., q)$, the random vectors $Y_i = (Y_{i1}, ..., Y_{iq})$ and the random matrix $Y = (Y_1^t, ..., Y_n^t)^t$. Let x_{ij} $(i = 1, ..., n; j = 1, ..., p)$ be real numbers, $x_i = (x_{i1}, ..., x_{ip})$ and $x = (x_1^t, ..., x_n^t)^t$. Moreover, let β be a vector in $\mathbb{R}^{s\times 1}$, $\mathbf{x}_i$ a $q \times s$ matrix created from x_i, $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$, and $h : z = (z_1, ..., z_q) \mapsto (h_1(z), ..., h_q(z))$ an invertible function. Then, $(Y, x, \beta, \mathbf{x}, h)$ is called a generalized linear model, if (G1) and (G2) hold:

(G1) *Distribution assumption:*

(a) There is a simple, q-parametric exponential family with scale parameter $(\mathbb{P}_{\theta,\psi})_{\theta\in\Theta,\psi\in\Psi}$ and one element $\psi \in \Psi$ with the property: For all $i = 1, ..., n$, the distribution of Y_i belongs to $(\mathbb{P}_{\theta,\psi})_{\theta\in\Theta}$.

(b) $Y_1, ..., Y_n$ are independent.

(G2) *Structure assumption:*

The expectation vector $\mu_i = \mathbb{E}(Y_i)$ and the linear predictor $\eta_i = (\mathbf{x}_i\beta)^t$ are connected by h, that is, $\mu_i = h(\eta_i)$.

In applications, n is the sample size while x_i and Y_i represent the values of the covariates and the endogenous characteristics corresponding to the i-th sample unit.

B.2 GLMs with stochastic covariates

Consider Y, β and h as in B.1. Let X_{ij} $(i = 1, ..., n; j = 1, ..., p)$ be random variables, $X_i = (X_{i1}, ..., X_{ip})$ and $X = (X_1^t, ..., X_n^t)^t$. Moreover, let $\mathbf{X}_i$ be a $q \times s$ matrix created from X_i, $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_n)$. We call $(Y, X, \beta, \mathbf{X}, h)$ a GLM with stochastic covariates, if:

(G1) *Distribution assumption:*

(a) There is a simple, q-parametric exponential family with scale parameter $(\mathbb{P}_{\theta,\psi})_{\theta\in\Theta,\psi\in\Psi}$ and one element $\psi \in \Psi$ with the property: For all $i = 1, ..., n$ and all possible realizations x of X, the conditional distribution of Y_i given $X = x$ belongs to $(\mathbb{P}_{\theta,\psi})_{\theta\in\Theta}$.

(b) $Y_1, ..., Y_n$ are independent given $X = x$ (for any value x of X).

(G2) *Structure assumption:*

The conditional expectation $\mu_i = \mathbb{E}(Y_i|X)$ and $\eta_i = (\mathbf{X}_i\beta)^t$ are connected by $\mu_i = h(\eta_i)$.

C Fisher scoring in GLM

Fisher scoring is an iterative method to compute maximum likelihood estimates. Notice, in (G1) from B.1 respectively B.2 the set of scale parameters Ψ appears. We describe Fisher scoring only for the case $\Psi = \{1\}$, because this case is relevant in this article.

C.1 Fisher scoring in GLMs with deterministic covariates

Let $(Y, x, \beta, \mathbf{x}, h)$ be a GLM and $y = (y_1^t, ..., y_n^t)^t \in \mathbb{R}^{n \times q}$ an observed value of Y. According to (G1), we need to maximize $l(\beta) = l(\beta, y) = \sum_{i=1}^n l_i(\beta)$ in β where $l_i(\beta) = l_i(\beta, y) = \theta_i y_i^t - b(\theta_i)$. To maximize l, the Fisher scoring algorithm generates a sequence of estimates $(\beta_\nu)_{\nu \in \mathbb{N}_0}$ as follows: When an estimate β_ν is available from the preceding iteration, the next estimate is computed by

$$\beta_{\nu+1} = \beta_\nu + F^{-1}(\beta_\nu) \cdot s(\beta_\nu). \tag{25}$$

Here, $s(\beta) = s(y, \beta) = (l'(\beta))^t$ is called score function, where $l'(\beta) \in \mathbb{R}^{1 \times s}$ denotes the Jacobi matrix of l at β, and $F(\beta) = \mathbb{E}[-\frac{d^2}{d\beta^2}\, l(Y, \beta)] = Var(s(Y, \beta))$ is the Fisher matrix. Define the partial score functions $s_i(\beta) = s_i(y, \beta) = (l_i'(\beta))^t$ and the partial Fisher matrices $F_i(\beta) = Var(s_i(Y, \beta))$. We have $s(\beta) = \sum_{i=1}^n s_i(\beta)$ and can show by standard calculations that $s_i(\beta) = \mathbf{x}_i^t \cdot D_i(\beta) \cdot [\Sigma_i(\beta)]^{-1} \cdot (y_i - \mu_i(\beta))^t$ with

$$D_i(\beta) = \left[h'((\mathbf{x}_i\beta)^t)\right]^t, \quad \Sigma_i(\beta) = Var_\beta(Y_i), \quad \mu_i(\beta) = h((\mathbf{x}_i\beta)^t),$$

where $h'(z)$ represents the Jacobi matrix $(D_j h_i(z))_{i,j=1,...,q}$. Moreover, $F(\beta) = \sum_{i=1}^n F_i(\beta)$ and $F_i(\beta) = \mathbf{x}_i^t \cdot W_i(\beta) \cdot \mathbf{x}_i$ hold, where $W_i(\beta) = D_i(\beta) \left[\Sigma_i(\beta)\right]^{-1} D_i(\beta)^t$.

We notice that the number of computations for Fisher scoring can be reduced when the number of different covariate levels is smaller than the number of rows of x: Let $g \leq n$ be the number of different rows of x, i.e., we have g covariate levels. We introduce the sets $(r = 1, ..., g)$

$$I_r = \{i \in \{1, ..., n\} \ : \ \text{sample unit } i \text{ possesses covariate level } r\},$$

define n_r to be the number of elements in I_r and assume $i_1 \in I_1$,...,$i_g \in I_g$. We remark that all units with the same covariate level have identical values for $\mu_i(\beta)$, i.e, $\mu_i(\beta) = \mu_j(\beta)$ for $i, j \in I_r$ $(r = 1, ..., g)$. An analog statement holds for $D_i(\beta)$, $\Sigma_i(\beta)$, $W_i(\beta)$ and $F_i(\beta)$. For this reason, we can conclude

$$F(\beta) = \sum_{r=1}^g n_r \cdot F_{i_r}(\beta) \text{ and } s(\beta) = \sum_{r=1}^g \mathbf{x}_{i_r}^t D_{i_r}(\beta) \left[\Sigma_{i_r}(\beta)\right]^{-1} n_r \left[\left(\frac{1}{n_r} \sum_{i \in I_r} y_i^t\right) - \mu_{i_r}(\beta)^t\right].$$

Hence, to obtain $F(\beta)$ and $s(\beta)$, we have to sum up each g terms. When g is considerably smaller than n, the effort to calculate $F(\beta)$ and $s(\beta)$ decreases significantly.

C.2 Fisher scoring in GLMs with stochastic covariates

Consider a GLM with stochastic covariates $(Y, X, \beta, \mathbf{X}, h)$ and assume y and x are observed realizations of Y and X respectively. As usual, let $f_{Y_i|X}(\cdot \,|\, x)$ denotes the density of Y_i given $X = x$. We have to maximize the function $\beta \mapsto \prod_{i=1}^n f_{Y_i|X}(y_i|x)$. However, this function is the likelihood function corresponding to a GLM with deterministic covariates. Thus, we can apply C.1.

References

[1] Fahrmeir L. / Tutz G. (2010): Multivariate statistical modelling based on generalized linear models. Springer.

[2] Gentle J.E. (1998): Random Number Generation and Monte Carlo Methods. Springer.

[3] Greenberg B.G. / Abul-Ela A.A. / Simmons W.R. / Horvitz D.G. (1969): The Unrelated Question Randomized Response Model: Theoretical Framework. Journal of the American Statistical Association 64, 520-539.

[4] Groenitz, H. (2012): A New Privacy-Protecting Survey Design for Multichotomous Sensitive Variables. Metrika, DOI: 10.1007/s00184-012-0406-8.

[5] Kuk A.Y.C. (1990): Asking Sensitive Questions Indirectly. Biometrika 77, 436-438.

[6] Maddala G.S. (1983): Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press.

[7] Scheers N.J. / Dayton C.M. (1987): Improved Estimation of Academic Cheating Behavior Using the Randomized Response Technique. Research in Higher Education 26, 61-69.

[8] Scheers N.J. / Dayton C.M. (1988): Covariate Randomized Response Models. Journal of the American Statistical Association 83, 969-974.

[9] Van der Heijden P.G.M. / Van Gils G. (1996): Some logistic regression models for randomized response data. Proceedings of the 11th International Workshop on Statistical Modelling (Orvieto, Italy 15-19 July, 1996), 341-348.

[10] Van den Hout A. / Van der Heijden P.G.M. / Gilchrist R. (2007): The logistic regression model with response variables subject to randomized response. Computational Statistics & Data Analysis 51, 6060-6069.

[11] Warner S.L. (1965): Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association 60, 63-69.

```
function [beta, Iter, SE,V_beta, p_beta, fit]=...
    fisherscore1(X,Y,model,C0,BETA0,epsilon)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Supplemental material for the manuscript
% Groenitz, H.: A Covariate Nonrandomized Response Model for
% Multicategorical Sensitive Variables.

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%



%This program can be applied to estimate parameters (a) in logistic
%regression models and (b) according to LR-DM estimation

% I N P U T

% X: design matrix. The number of rows in X is the number of different
% covariate levels, the number of columns is the number of covariates.
% Y: response matrix with q+1 columns. The entry Y_ij represents the
% absolute frequency of category j among the units having the i-th
% covariate level.
% model: When intending to analyze an ordinary logistic regression model,
% type 'logreg'. When considering the diagonal model with covariates and
% intending to conduct a LR-DM estimation type 'diagcov'.
% C0: (q+1) x (q+1) design matrix in the diagonal model, every row is a
% left-cyclic shift of the row above (for model 'logreg' an arbitrary
% (q+1) x (q+1) matrix can be typed for C0).
% BETA0: starting values for Fisher scoring algorithm
% epsilon:  accuracy of calculation

% O U T P U T

% beta: vector of estimated parameters (maximum likelihood estimate, MLE)
% Iter: number of iterations of Fisher scoring algorithm
% SE: estimated standard errors for the estimation
% V_beta: estimated variance matrix of the estimator
% p_beta: p-values for the tests with H_0: beta_i=0
% fit=[chi2, pchi2, dev, pdev, df] where
% chi2: value of the test statistic for the chi^2-goodness-of-fit test
% pchi2: p-value for chi^2-goodness-of-fit test
% dev: value of the test statistic for the deviance test (this is another
% well-known goodness-of-fit test, cf. ``Multivariate Statistical Modelling
% Based on Generalized Linear Models'' by Fahrmeir and Tutz (2010),
% Springer, page 50)
% pdev: p-value for deviance test
% df: degrees of freedom for chi^2 / deviance test

% E X A M P L E 1 (estimation in logistic regression models)
% The data of this example are taken from an example in the book "Multivariate
% statistische Verfahren'' by Fahrmeir et al. (1996), de Gruyter, page
% 263 / 267 where the sales of gasoline stations are investigated.

% The rows of the following matrix X represent the observed covariate
% levels
% x1=ones(1,12)';x2=[ones(1,6) -1 -1 -1 -1 -1 -1]';
% x3=[1 1 1  -1 -1 -1  1 1 1  -1 -1 -1]';x4=[1 0 -1 1 0 -1 1 0 -1 1 0 -1]';
% x5=[0 1 -1 0 1 -1 0 1 -1 0 1 -1]'; X=[x1 x2 x3 x4 x5];
```

```
% Each row of the following matrix Y contains the absolute frequencies of
% the categories 1 (low sales), 2 (medium sales), 3 (large sales) for the
% corresponding covariate level
% y1=[2 2 3 65 63 48 4 2 5  38 16 179]';y2=[3 0 4 32 24 12 4 0 12 19 7  55]';
% y3=[0 0 1 20  4  6 7 1  4 27 2  29]'; Y=[y1 y2 y3];

% Set
% beta0=zeros(10,1); para=eye(10);
% then the command
% [beta, Iter, SE,V_beta, p_beta, fit]=fisherscore1(X,Y,'logreg',para,beta0,10^-8)
% delivers among others the MLE beta:
% 1.2209 0.3735 -0.5320 -0.9716 0.6174 0.8744 0.3542 0.0978 -0.7246 0.5615

% EXAMPLE 2 (Diagonal model with covariates, LR-DM estimation)

% We introduce the following quantities
% X=[1 1 1 1 1; 1 2 3 4 5]';
% Y=[35 16 30; 27 18 35; 20 22 38; 16 27 36; 15 33 33];
% C0=[2/3 1/6 1/6; 1/6 1/6 2/3; 1/6 2/3 1/6];BETA0=[0 0 0 0; 1 -1 1 -1];
% That is, we have two covariates, and the available covariate levels are
% (1,1),...,(1,5). E.g., for covariate level (1,1), we have 35 respondents
% giving diagonal model answer 1, 16 respondents giving answer 2 and 30
% respondents giving answer 3. The command
% [beta, Iter, SE,V_beta, p_beta, fit]=fisherscore1(X,Y,'diagcov',C0,BETA0,10^-8)
% returns among others the estimate beta equal to
% 3.5691  -1.2722  2.5304  -0.5052

%-------------------------------------------------------------------
q=length(Y(1,:))-1; R=q+1; n=length(X(:,1)); p=length(X(1,:)); nn=sum(Y,2);
if min(nn)==0
   error('n_i equals 0 for some i; Remove corresponding rows in X and Y.')
end

%----- Def. of functions-------------------------------------------
Q =@(z)sum(exp([0 z])); %z row vector
Jh=@(z)( diag(exp(z)*Q(z)) - exp(z')*exp(z)) /(Q(z))^2;
%          D h1
% Jh = [     .      ]   Jacobi matrix of h
%          D hq
h =@(z)exp(z)/Q(z);

CC=C0(1:q,1:q);
for j=1:q
   CC(:,j)=CC(:,j)-C0(1:q,R);
end
m =@(z)h(z) * CC' + C0(1:q,R)';
function M=Jm(z,CC,q,Jh) %Jacobi matrix of m % "nested function"
M=zeros(q); JJ=feval(Jh,z);
for l=1:q
   M=M + CC(:,l)*JJ(l,:);
end
end
%-------------------------------------------------------------------

if strcmp(model,'logreg')   %compares strings
% Here, the case of a logistic regression model is studied.

beta0=BETA0;
```

```
for j=1:n
    Y(j,:)=Y(j,:)/nn(j);
end
Y=Y(:,1:q);
b0=beta0;

F=0;score=0;
for i=1:n
    X_i     =X(i,:);
    for j=2:q
        X_i=blkdiag(X_i,X(i,:)); %block diagonal matrix
    end
    P_i     =(X_i*b0)' ; %predictor
    D_i     =Jh(P_i)';
    mu_i    =h(P_i);
    Sigma_i =(diag(mu_i)-mu_i' * mu_i)/nn(i);
    W_i     =D_i * inv(Sigma_i) * D_i';

    F=F + X_i' * W_i * X_i;
    score=score + X_i'*D_i * inv(Sigma_i) * (Y(i,:)'-mu_i');
end
b1=b0 + F\score; %A^-1 * b : A\b

Iter=1;

while norm(b1-b0)/norm(b0)>epsilon
    Iter=Iter+1;
    b0=b1;

    F=0;score=0;
    for i=1:n
    X_i     =X(i,:);
    for j=2:q
        X_i=blkdiag(X_i,X(i,:));
    end
    P_i     =(X_i*b0)';  %predictor
    D_i     =Jh(P_i)';
    mu_i    =h(P_i);
    Sigma_i =(diag(mu_i)-mu_i' * mu_i)/nn(i);
    W_i     =D_i * inv(Sigma_i) * D_i';

    F=F + X_i' * W_i * X_i;
    score=score + X_i'*D_i * inv(Sigma_i) * (Y(i,:)'-mu_i');
    end

    b1=b0 + F\score; % A^-1 * b: A\b
end

beta=b1;

% Standard errors, testing H_0: beta_i=0, goodness-of-fit tests (chi^2 /
% deviance)

chi2=zeros(n,1); dev=zeros(n,1); F=0;
for i=1:n
    X_i     =X(i,:);
    for j=2:q
        X_i=blkdiag(X_i,X(i,:));
```

```
    end
    P_i     =(X_i*beta)';  %beta: MLE
    D_i     =Jh(P_i)';
    mu_i    =h(P_i);
    Sigma_i =(diag(mu_i)-mu_i' * mu_i)/nn(i);
    W_i     =D_i * inv(Sigma_i) * D_i';

    %for Fisher matrix at the MLE beta
    F=F + X_i' * W_i * X_i;

    %for chi2-goodness-of-fit test
    chi2(i)=(Y(i,:)-mu_i)* inv(Sigma_i) *(Y(i,:)-mu_i)';

    % for deviance; mnpdf(X,PROB) X and PROB  1-by-k vectors, where k is the
    % number of multinomial categories
    Z_i=round( [Y(i,:) 1-sum(Y(i,:))]*nn(i) ); %abs. frequencies

    L1=mnpdf(Z_i, [mu_i 1-sum(mu_i)]);    l1=log(L1);
    L2=mnpdf(Z_i, Z_i/nn(i));  l2=log(L2);
    dev(i)=l1-l2;
end

%Estimated standard errors for the components of the MLE
SE=sqrt(diag(inv(F)));

%Estimated variance matrix for the MLE
V_beta=inv(F);

%Testing H_0: beta_i=0 (t-statistics; p-values)
T=beta./SE; p_beta=2*(1-normcdf( abs(T) ) );

% goodness-of-fit
CHI2=sum(chi2); DEV=-2*sum(dev);
df=n*q-p*q; % degrees of freedom
%p-values
pCHI2=1-chi2cdf(CHI2,df); pDEV=1-chi2cdf(DEV,df);
fit=[CHI2,pCHI2,DEV,pDEV,df];

end

% ---------------------------------------------------------

if strcmp(model,'diagcov')
% Case of diagonal model with covariates, LR-DM estimation is conducted.

YY=Y; %for later calculation of the log-Likelihood
for j=1:n
    Y(j,:)=Y(j,:)/nn(j);
end
Y=Y(:,1:q);

E=zeros(length(BETA0(:,1)),p*q+1);

for jj=1:length(BETA0(:,1))
    beta0=BETA0(jj,:)';

    b1=beta0;
    cond=inf;Iter=1;
    while cond>epsilon
```

```
        Iter=Iter+1;
        b0=b1;

        F=0;score=0;
        for i=1:n
        X_i     =X(i,:);
        for j=2:q
         X_i=blkdiag(X_i,X(i,:));
        end
        P_i     =(X_i*b0)';
        D_i     =Jm(P_i,CC,q,Jh)';
        mu_i    =m(P_i);
        Sigma_i =(diag(mu_i)-mu_i' * mu_i)/nn(i);
        W_i     =D_i * inv(Sigma_i) * D_i';

        F=F + X_i' * W_i * X_i;
        score=score + X_i'*D_i * inv(Sigma_i) * (Y(i,:)'-mu_i');
        end

        b1=b0 + F\score; %A^-1 * b = A\b
        cond=norm(b1-b0)/norm(b0);

        %To avoid endless loops
        if Iter > 1000
           b1=ones(p*q,1)*NaN;
           cond=0;
        end

    end %endwhile

    beta=b1;

    %Plausibility check

    if sum(isnan(beta))==0 && sum(isinf(beta))==0 && rcond(F)<10^-15
       beta=ones(p*q,1)*NaN;
    end
    %now a beta for this starting value is available
E(jj,1:p*q)=beta';
mu=zeros(n,q+1);
for i=1:n
    eta_i=zeros(1,q);
      for j=1:q
              eta_i(j)=X(i,:)*beta( (j-1)*p+1: j*p);
      end
    mu(i,1:q)=m(eta_i);
end
mu(:,q+1)=1-sum(mu(:,1:q),2);
E(jj,p*q+1)=sum(sum(YY.*log(mu)));
% value of the log-likelihood (Y: frequencies of the answers)
end  %end jj-loop

% Which starting value leads to the largest likelihood?
% The max function ignores NaNs. max([0 1 Nan])=1; max([NaN NaN])=NaN;

M=max(E(:,p*q+1));

if isnan(M)==1
    beta=ones(p*q,1)*NaN;
```

```
else
    ind=find(E(:,p*q+1)==M);
    ind=ind(1);
    beta=E(ind,1:p*q)';
end

% Standard errors, testing H_0: beta_i=0, goodness-of-fit tests (chi^2 /
% deviance)

chi2=zeros(n,1); dev=zeros(n,1); F=0;
for i=1:n
    X_i     =X(i,:);
    for j=2:q
        X_i=blkdiag(X_i,X(i,:));
    end
    P_i     =(X_i*beta)';
    D_i     =Jm(P_i,CC,q,Jh)';
    mu_i    =m(P_i);
    Sigma_i =(diag(mu_i)-mu_i' * mu_i)/nn(i);
    W_i     =D_i * inv(Sigma_i) * D_i';

    %for Fisher matrix at the MLE beta
    F=F + X_i' * W_i * X_i;

    %for chi2-goodness-of-fit test
    chi2(i)=(Y(i,:)-mu_i)* inv(Sigma_i) *(Y(i,:)-mu_i)';

    % for deviance test; mnpdf(X,PROB) X and PROB  1-by-k vectors, where k is the
    % number of multinomial categories
    Z_i=round( [Y(i,:) 1-sum(Y(i,:))]*nn(i) ); %abs. frequencies

    L1=mnpdf(Z_i, [mu_i 1-sum(mu_i)]);    l1=log(L1);
    L2=mnpdf(Z_i, Z_i/nn(i));  l2=log(L2);
    dev(i)=l1-l2;
end

%Estimated standard errors for the components of the MLE
SE=sqrt(diag(inv(F)));

%Estimated variance matrix for the MLE
V_beta=inv(F);

%Testing H_0: beta_i=0 (t-statistics, p-values)
T=beta./SE; p_beta=2*(1-normcdf( abs(T) ) );

%for goodness-of-fit tests
CHI2=sum(chi2); DEV=-2*sum(dev);
df=n*q-p*q; % degrees of freedom
%p-values
pCHI2=1-chi2cdf(CHI2,df); pDEV=1-chi2cdf(DEV,df);
fit=[CHI2,pCHI2,DEV,pDEV,df];
end

end
```